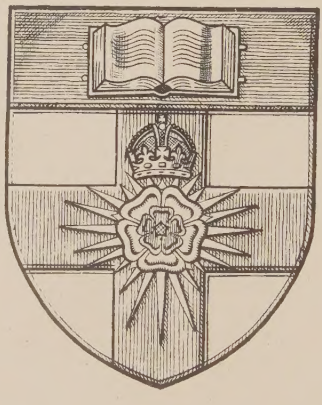
P.S. [illegible] Depository

SHL
SHL
WITHDRAWN
W



UNIVERSITY OF LONDON
LIBRARY

Digitized by the Internet Archive
in 2026

https://archive.org/details/british-journal-of-psychology_1951_4

THE BRITISH JOURNAL OF PSYCHOLOGY

Statistical Section

EDITED BY
CYRIL BURT AND GODFREY THOMSON

WITH THE ASSISTANCE OF
CHARLOTTE BANKS AND ARTHUR SUMMERFIELD
AND THE FOLLOWING EDITORIAL BOARD

A. C. AITKEN
M. S. BARTLETT
W. G. EMMETT
M. G. KENDALL
D. N. LAWLEY
E. S. PEARSON
L. S. PENROSE
J. FRASER ROBERTS
A. RODGER
B. BABINGTON SMITH
W. STEPHENSON
P. E. VERNON

Managing Editor: H. G. MAULE

Volume IV
1951

UNIVERSITY OF LONDON PRESS LTD.
WARWICK SQUARE, LONDON, E.C.4.

BIBL.
LONDIN.
UNIV.

Vol. IV
Part I

The British Journal of Psychology,
Statistical Section

March,
1951

A FURTHER NOTE ON TESTS OF SIGNIFICANCE IN FACTOR ANALYSIS

By M. S. BARTLETT

Department of Mathematics, University of Manchester

In my previous paper (1) I was obliged to retain in the proposed significance test for principal components analysis an unresolved doubt on the appropriate number of degrees of freedom to be assigned to the reduced χ^2, after elimination of one or more significant components. To quote from Part III of my paper (top of p. 84) :

" One point that remains in some doubt pending a more detailed examination is whether the reduction in degrees of freedom that ensues from the individual standardization of the tests is automatically felt in the residual factor components (an assumption implicit in the proposed test) or is mainly absorbed by the larger roots."

A ' more detailed examination ' of a particular case, while confirming this extra complication introduced by the standardization, has also helped to resolve this point, and is therefore summarized below. The algebraic details, which are of some complexity, are omitted, but it is intended to record them elsewhere (2).

Consider a sample correlation matrix obtained from p tests, between all of which there is the same true correlation ρ. The first true latent root is then $1 + (p - 1)\rho$; the others are all equal at $1 - \rho$ and thus homogeneous. The quantity χ^2 defined in equation (3) of my previous paper, with $k = 1$ (i.e., with the first component removed), was examined in the case $\rho \neq 0$, as n increases. Its limiting expected value was found to be

$$\tfrac{1}{2}(p - 2)(p + 1) - \frac{(p - 1)(p - 2)}{p}(1 - \rho)^2 ,$$

which will be seen to be greater than $\tfrac{1}{2}(p - 2)(p - 1)$, and to rise to $\tfrac{1}{2}(p - 2)(p + 1)$ as $\rho \to 1$. For example, in the particular case $p = 3$, the expected value is not 1, as originally suggested, but rises to 2 as $\rho \to 1$.

The conclusion is thus that the reduction in the total number of degrees of freedom due to standardization is only *partially* felt in the smaller roots and disappears entirely when the components eliminated are large enough. That this rather peculiar result is essentially associated with standardization may be demonstrated by the following further case. The case $p = 3$ was re-examined for a sample variance-covariance matrix from the same population, with constant true variance for each test standardized to unity for the *mean* sample variance. In these circumstances the elimination of the largest component gives rise to a reduced χ^2 which should be valid under the same conditions as in external canonical analysis (namely, for a real effect which is sufficiently well-determined, i.e., at least when n is large). This was confirmed, the limiting expected value of χ^2 as n increases being 2 *for any value of* $\rho > 0$.

It is evident in the standardized case that the χ^2 approximation cannot strictly be valid except in the further limit of $\rho \to 1$, when the remaining components

contribute negligibly to the total variance. That it *is* valid in this limit hardly follows merely from the integral value for the expected value ; it may, however, be checked from its asymptotic form before averaging (this verification has been completed for $p = 3$ and 4).

The general practical conclusion to be drawn from these further detailed investigations is thus that the effect of standardization cannot (unfortunately) be very exactly allowed for in the reduced χ^2 after one or more significant components have been eliminated, but that, as its effect tends to be lost with the elimination of large components, *it is safer to allocate as the degrees of freedom of the reduced* χ^2 *the numbers appropriate to a test on the unstandardized matrix*, viz., $\frac{1}{2}(p-k-1)(p-k+2)$. The only place in the numerical examples previously discussed (1) where the conclusions on significance thereby require amendment is in the table at the bottom of p. 79, where, since the effective number of degrees of freedom in the second line can now rise to 5, homogeneity of the last three roots is not disproved.

Since the best multiplying coefficient for the total χ^2 on an unstandardized matrix is not $n - \frac{1}{6}(2p + 5)$ but $n - \frac{1}{6}(2p + 1 + 2/p)$, it might be argued that the latter factor, and not the former, should now be adapted for the best approximation for the reduced χ^2, but clearly, in view of the more serious complications noted, such refinement is not justified.

The above results, and consequent amendment of the number of degrees of freedom, form rather a striking example of the inadequacy by itself (with these non-linear analyses) of the arguments on degrees of freedom based merely on numbers of constants fitted. In the χ^2 analyses for the standardized and unstandardized cases, say for $p = 3$, we start with 3 and 5 degrees of freedom respectively, eliminate the largest component by the same arithmetical procedure, and yet we can still finish with the same remaining number of degrees of freedom, namely 2.

REFERENCES

1. Bartlett, M. S. (1950). ' Tests of significance in factor analysis.' *Brit. J. Psych., Stat. Sect.*, III, 77–85.
2. Bartlett, M. S. (1951). ' The effect of standardization on a χ^2 approximation in factor analysis.' *Biometrika* (in the press).

Vol. IV
Part I

The British Journal of Psychology,
Statistical Section

March,
1951

BILINGUALISM AND VERBAL INTELLIGENCE

By W. R. JONES

University College of North Wales, Bangor

and

W. A. C. STEWART

University College of North Staffordshire [1]

I. *Introduction.* II. *Purpose and Plan of the Investigation.* III. *Statistical Procedure.* IV. *Statistical Analysis.* V. *Summary and Conclusions.*

I. INTRODUCTION

It will be helpful first of all to summarize previous investigations in Wales that have dealt with the relation between bilingualism and intelligence as measured by means of verbal tests. The earliest are those of D. J. Saer and Frank Smith (1, 2, 3, 4). The main part of Saer's investigation showed the inferiority of rural bilinguists to rural monoglots, when tested by the Stanford Revision of the Binet scale and by Burt's version ; no such difference was found in the intelligence of bilingual and monoglot children in urban areas. Group verbal intelligence tests were also administered by Saer to University students ; he concluded that bilingual students from rural areas were inferior to monoglots, whereas no such difference was observed between those from urban areas. His findings were supplemented by the work of Frank Smith. Four verbal tests (free composition, word building, mutilated fables, and an analogies test) were administered three times at intervals of one year to monoglot and bilingual children in four urban schools. When the results were averaged, the monoglots' superiority was established in each of the tests.

In more recent investigations (5, 6) the problem has been approached from a different angle. Ethel M. Barke administered the Pintner Non-Language Test and Burt's Northumberland Standardized Test to bilingual children in a Welsh mining district and monoglot children in an English-speaking mining district. She found that the monoglots were superior in the verbal test and the bilinguists in the non-verbal test. Ethel M. Barke and D. E. Parry Williams collaborated in a further investigation in which four tests were used : (i) the Pintner Non-Language Test, (ii) Burt's Northumberland Standardized Test, (iii) the 1932 Mental Survey Test of the Scottish Council for Research in Education, and (iv) the Thorndike Test of Word Knowledge—Form A (in Welsh) and Form B (in English). These were given to three groups, each drawn from two schools, and aged $10\frac{1}{2}$ to $11\frac{1}{2}$ years. This further investigation confirmed the results of the previous enquiry : the difference between the bilingual and monoglot groups, when measured by an intelligence test in which no language was used, was insignificant, but the bilinguists were distinctly inferior to the monoglots in a verbal test administered in their second language,

[1] This work was undertaken when the authors were members of the Education Department of the University College of South Wales and Monmouthshire, Cardiff, on the advice of Professor Dame Olive Wheeler, who helped to shape the experimental design and offered guidance on many points. The authors collaborated on the practical testing and compiling of results. The statistical analysis has been conceived and carried through by Mr. Jones. He wishes to acknowledge the generous help and advice he has had from Professor Sir Godfrey Thomson, Dr. Lawley, and Professor Sir Cyril Burt in preparing the work for publication.

English ; furthermore, when the two groups were assessed by an intelligence test given to each in its mother-tongue, the superiority of the monoglots was still more marked ; it was also observed that the bilinguists did not have a vocabulary equal to that of the monoglots in either language.

Results hitherto obtained in Wales may, therefore, be briefly summarized as follows. (1) Bilingual and monoglot groups differ little in non-verbal intelligence tests. (2) Monoglot groups are usually superior to bilingual groups in verbal tests.

That the results are by no means conclusive is suggested by similar investigations elsewhere (7). Their exact interpretation is often rendered difficult by a number of complicating factors both in the experimental situation itself and in the subsequent treatment of data. The following defects may be mentioned. (i) The statistical significance of actual differences between group means is not indicated in any of the researches reported in the preceding paragraphs. (ii) The test material is often unsuitable in validity or reliability : this applies particularly to the work of Saer and Smith. (iii) Translation of a verbal test into Welsh does not necessarily ensure equivalence of difficulty in the Welsh and English versions ; and this may vitiate comparisons between monoglot and bilingual groups. (iv) No language questionnaire is used to obtain an objective estimate of the pupils' bilingual background. Such an assessment is especially important in Wales, because of the complexity of the language position, and seems essential to a correct interpretation of any test results.

II. PURPOSE AND PLAN OF THE INVESTIGATION

Purpose. The main purpose of the present enquiry was to obtain further information about the comparative performance of monoglots and bilinguists in verbal intelligence tests.

Material. (*a*) *The Tests.* The tests used were as follows :

1. A Scale of Non-verbal Mental Ability.[1]
2. The Cotswold Mental Ability Test, Series I.[2]

The first test is perceptual : no verbal responses are required ; the pupils merely underline the appropriate design in the test booklet. For bilinguists the instructions were translated into Welsh.

The second test is verbal ; it calls for ability in reading, but not in writing, since the pupils are required only to underline. This was adapted into Welsh ; and care was taken to ensure that the original version and the Welsh were as far as possible of equal difficulty. The final form of the Welsh version was decided after a preliminary trial with about one hundred bilingual children.

(*b*) *The Testees.* For the main investigation, 518 bilingual children were obtained from three Welsh-speaking areas, and 326 monoglot children from an English-speaking area. All areas were predominantly rural ; and answers to a simple sociological questionnaire showed that the environment of both groups was similar. The children tested consisted of all those present in the schools on the day of examination whose ages fell between $10\frac{1}{2}$ and $11\frac{1}{2}$.

Each bilingual pupil was required to answer a brief questionnaire about the linguistic background of his home. From the replies it appeared that this group was composed chiefly of children who came from homes where Welsh was always spoken, and whose knowledge of English had been acquired outside the home, mainly in school. However, in a few cases one parent spoke English, in others Welsh-speaking parents used both languages. Hence a small number were acquainted with Welsh and English at home.

Experimental Design. The plan of the present investigation was as follows :

(*a*) To test two similar groups, monoglots and bilinguists respectively, by a non-verbal and a verbal test of intelligence, administered in Welsh to the bilingual group and in English to the monoglot group.

(*b*) To equate the two groups in the non-verbal intelligence test.

(*c*) To determine whether any significant difference remained between them in the verbal test after allowance had been made for any initial difference in the non-verbal test.

[1] J. W. Jenkins, The National Foundation for Educational Research in England and Wales.

[2] J. W. Jenkins. Glasgow : Robert Gibson and Sons, Ltd.

Certain features in the plan call for comment. First, it assumes that the non-verbal test measures the same factor of general intelligence as the verbal. Secondly, it does not necessarily assume that a non-verbal test measures nothing but intelligence *plus* a factor specific to itself. Previous investigations suggest the presence of a spatial factor in tests constructed from abstract diagrams, pictures, and other non-verbal material (8, 9, 10, 11). Emmett's work, however, indicates that the Jenkins non-verbal test is free from the spatial factor, and practically free from the verbal factor ; it also has a high saturation with the general factor (11, p. 8). For these reasons, it may be considered eminently suitable for the present experiments. Thirdly, the scheme assumes that, in addition to the general factor, verbal intelligence tests contain a verbal factor. The various researches summarized by Burt in his recent review of 'the results of factor analysis' (16, pp. 183f.) afford ample evidence on this point ; and it is fairly safe to conclude that all intelligence tests that depend on the manipulation of words involve a verbal factor. Fourthly, opinions apparently differ on the question how far non-verbal tests are free from the verbal factor. Some assume that this is so (12, p. 3) ; others hold that verbal processes may enter into the manipulation even of non-verbal material (13, pp. 18, 20, 21, and 183). Colvin and Allen (14) maintain that " it does not seem to follow necessarily when a test is employed in which no words are used, that for this reason linguistic ability is not brought into play, particularly when rational processes are involved." For the purpose of the present investigation, however, it is sufficient to observe that the non-verbal test is as non-verbal as anything can be, and that its use at any rate reduces the influence of the verbal factor to a minimum.

III. STATISTICAL PROCEDURE

It is necessary at this stage to consider which statistical procedure may be most suitably applied to the present enquiry. Two methods of equating groups have been widely practised in educational research :

(i) *The Method of Pairing Observations.* In the present experiment, pairing may be effected by arranging both the monoglot and bilingual group in the order of their scores in the non-verbal intelligence test and then eliminating all the pupils in one group, who cannot be paired with an equivalent pupil in the other.

(ii) *The Method of Matched Groups.* Alternatively, the monoglot and bilingual groups may be arranged, without any pairing of individuals as such, so that in the non-verbal intelligence test their respective means, standard deviations, and distributions are comparable. The use of either paired or matched groups involves a depletion in the number of pupils available, and a consequent loss of information.

(iii) *The Analysis of Covariance.* There is a third method whereby an equivalence in non-verbal intelligence and an increase in precision can be secured, namely, to analyse the covariances (15). It has the further advantage that it involves no diminution in numbers.

IV. STATISTICAL ANALYSIS

The original data, in summary form, are given in the following table.

TABLE I. SCORES OF MONOGLOT AND BILINGUAL GROUPS : (X = NON-VERBAL INTELLIGENCE, Y = VERBAL INTELLIGENCE)

Groups	n	SX	SY	SX^2	SY^2	SXY
Monoglot	326	10,626	14,607	449,816	811,491	577,033
Bilingual	518	12,850	17,829	443,222	807,423	558,631
Experimental totals	844	23,476	32,436	893,038	1,618,914	1,135,664

The first question is whether the two groups, monoglot and bilingual, differ significantly in respect of verbal intelligence. The test of significance is made in Table II.

TABLE II. COMPARISON OF MONOGLOT AND BILINGUAL GROUPS IN VERBAL INTELLIGENCE (Y)

Groups	n	Degrees of Freedom	Mean Score	Sum of Squares
Monoglot ..	326	325	44·81	156,998·83
Bilingual	518	517	34·42	193,768·10
Sum or Difference	844	842	10·39	350,766·93

Pooled Variance (s^2) : 350,766·93/842 = 416·59

Standard Error of Mean Difference ($s_{\bar{y}} = \sqrt{s^2(n_1 + n_2)/n_1 n_2}$) : 1·44

Critical Ratio : $t = 10{\cdot}39/1{\cdot}44 = 7{\cdot}22$

The results reveal a highly significant difference between the groups : the observed value of t is well beyond the 1 per cent. point for 842 degrees of freedom. It is possible, however, that this is partly due to a difference in non-verbal intelligence. The significance of such a difference may be tested in the same way. Accordingly the appropriate sums of squares have been again computed from the relevant summed values in Table I. The results are entered in Table III.

TABLE III. COMPARISON OF MONOGLOT AND BILINGUAL GROUPS IN NON-VERBAL INTELLIGENCE (X)

Groups	n	Degrees of Freedom	Mean Score	Sum of Squares
Monoglot ..	326	325	32·60	103,460·55
Bilingual	518	517	24·81	124,452·70
Sum or Difference	844	842	7·79	227,913·25

Pooled Variance (s^2) : 227,913·25/842 = 270·68

Standard Error of Mean Difference ($s_{\bar{x}} = \sqrt{s^2(n_1 + n_2)/n_1 n_2}$) : 1·16

Critical Ratio : $t = 7{\cdot}79/1{\cdot}16 = 6{\cdot}72$

TABLE IV. REGRESSION AND CORRELATION DATA IN MONOGLOT AND BILINGUAL GROUPS

Groups	Degrees of Freedom	Sums of Squares and Products			Correlation[1]	Coefficient[2]	Errors of Estimate	
		Sx^2	Sxy	Sy^2			Sum of Squares[3]	Degrees of Freedom
Monoglot	325	103,460·55	100,916·49	156,998·83	0·7918	0·9754	58,563·84	324
Bilingual	517	124,452·70	116,347·89	193,768·10	0·7492	0·9349	84,997·90	516
Sum ..							143,561·74	840
Average within groups	842	227,913·25	217,264·38	350,766·93	0·7684	0·9533	143,653·86	841

[1] Correlation = $\dfrac{Sxy}{\sqrt{(Sx^2)\,(Sy^2)}}$ [2] Regression = $\dfrac{Sxy}{Sx^2}$ [3] Sum of Squares = $Sy^2 - \dfrac{(Sxy)^2}{Sx^2}$

Since the observed value of t again lies well beyond the 1 per cent. point for 842 degrees of freedom, it is evident that there is also a highly significant difference between the two groups in non-verbal intelligence. We have now to determine whether the difference in verbal intelligence remains significant if the two groups are first equated in the non-verbal test. This may be done by an analysis of covariance. The results are shown in Table IV.

Two questions may be answered from the figures there given. First, do the monoglot and bilingual groups differ significantly, when verbal intelligence scores have been adjusted to a common non-verbal basis? Secondly, is there a significant difference between the group regression coefficients?

Before answering the first question, it is necessary to calculate the sum of squares of errors of estimate, i.e., the adjusted sum of squares, for the experiment *as a whole*. The number of degrees of freedom is one less than the number for the total sum. The figures obtained are shown in Table V.

TABLE V. TEST OF SIGNIFICANCE OF DIFFERENCE BETWEEN ADJUSTED MEANS IN VERBAL INTELLIGENCE

Source of Variation	Errors of Estimate		
	Degrees of Freedom	Sum of Squares	Mean Square
Total	842	145,322·26	
Average within groups	841	143,653·86	171
Between adjusted means	1	1,668·40	1,668

Adjusted means : monoglot group, 40·25 ; bilingual group, 37·29

$F = 1{,}668/171 = 9{\cdot}75$

The adjusted means have been obtained from the formula, $Y - bx$, where Y is the unadjusted group mean in verbal intelligence, x the deviation of the group mean in non-verbal intelligence from the general mean for the same variable, and b the average regression coefficient within groups as recorded in Table IV. The observed value of F clearly indicates that the difference between the two groups in verbal intelligence remains significant, though materially reduced.

The test of significance between regression coefficients is made in Table VI.

TABLE VI. TEST OF SIGNIFICANCE BETWEEN REGRESSION COEFFICIENTS FOR MONOGLOT AND BILINGUAL GROUPS

Source of Variation	Errors of Estimate		
	Degrees of Freedom	Sum of Squares	Mean Square
Average within groups	841	143,654	
Deviations from individual group regressions	840	143,562	171
Between regression coefficients	1	92	92

It is clear that the mean square for difference among the group regression coefficients is less than the mean square for deviations from individual group regressions. One may, therefore, conclude that the regression of verbal on non-verbal scores is fundamentally the same in both groups.

V. SUMMARY AND CONCLUSIONS

1. Two groups of children aged $10\frac{1}{2}$ to $11\frac{1}{2}$, consisting of 326 monoglot children from a predominantly English-speaking area of Wales, and 518 bilingual children from a predominantly Welsh-speaking area, both living in a predominantly rural environment, were tested with non-verbal and verbal tests of intelligence. The tests were administered in English to the monoglot group, and in Welsh to the bilingual group.

2. In the raw results a highly significant difference in favour of the monoglot group was found in both the verbal and the non-verbal tests of intelligence.

3. To discover whether the difference observed in favour of the monoglot group in the verbal intelligence test remained significant, when the two groups were equated for the non-verbal intelligence test, an analysis of covariance was carried out. It was found that, as a result of adjusting the verbal scores to a common non-verbal basis, the difference was substantially reduced, but still remained statistically significant. It was therefore concluded that the bilingual children were significantly inferior to the monoglot children, even after full allowance has been made for the initial difference in the non-verbal intelligence test.

4. The regression of verbal on non-verbal intelligence scores proved to be closely similar for both monoglot and bilingual groups.

REFERENCES

1. Saer, D. J. (1922). ' An inquiry into the effect of bilingualism upon the intelligence of young children.' *J. Exp. Ped.*, VI, 232–40 and 266–74.
2. Saer, D. J. (1923). ' The effect of bilingualism on intelligence.' *Brit. J. Psych.*, XIV, 25–38.
3. Smith, Frank (1923). ' Bilingualism and mental development.' *Ibid.*, XIII, 271–82.
4. Saer, D. J., Smith, F., and Hughes, J. (1924). *The Bilingual Problem : A Study based upon Experiments and Observations in Wales.* Wrexham : Hughes and Son.
5. Barke, Ethel M. (1933). ' A study of the comparative intelligence of children in certain bilingual and monoglot schools in South Wales.' *Brit. J. Educ. Psych.*, III, 237–50.
6. Barke, Ethel M., and Parry Williams, D. E. (1938). ' A further study of comparative intelligence of children in certain bilingual and monoglot schools in South Wales.' *Ibid.*, VIII, 63–77.
7. Davies, M., and Hughes, A. G. (1927). ' An investigation into the comparative intelligence and attainments of Jewish and non-Jewish school children.' *Brit. J. Psych.*, XVIII, 134–47.
8. El Koussy, A. H. (1935) : ' Investigation into factors in tests involving the visual perception of space.' *Brit. J. Psych. Mon. Suppl.*, XX, 89.
9. Mellone, Margaret A. (1944). ' A factorial study of picture tests for young children.' *Brit. J. Psych.*, XXXV, 9–16.
10. Drew, L. J. (1947). ' The measurement of technical ability.' *Occ. Psych.*, XXI, 34.
11. Emmett, W. G. (1949). ' Evidence of a space factor at 11 + and earlier.' *Brit. J. Psych., Stat. Sect.*, II, 3–16.
12. Cattell, R. B. (1936). *A Guide to Mental Testing.* London : Harrap & Co.
13. Bartlett, F. C. (1932). *Remembering : A Study in Experimental and Social Psychology.* Cambridge : University Press.
14. Colvin, Stephen S., and Allen, Richard D. (1923). ' Mental tests and linguistic ability.' *Journ. Educ. Psych.*, XIV, 1–20.
15. Snedecor, George W. (1946). *Statistical Methods* (4th ed.). Iowa : Iowa State Collegiate Press.
16. Burt, C. (1949). ' The structure of the mind : A review of the results of factor analysis.' *Brit. J. Educ. Psych.*, XIX, 100–11 and 176–99.

Vol. IV
Part I

The British Journal of Psychology,
Statistical Section

March,
1951

THE TRANSFORMATION OF A MATRIX OF NEGATIVE CORRELATIONS

By PATRICK SLATER

Bedford College, London

I. *Problem and Data.* II. *Definition of Negative Matrices.* III. *The Transformation Proposed.* IV. *Treatment of the Data.* V. *Interpretation of Results.* VI. *Alternative Treatments.* VII. *Summary.*

I. PROBLEM AND DATA

When two variables are both negatively correlated with a third, it is natural to expect that they will correlate positively with one another. Conversely, a correlation table in which all the variables tend to be negatively correlated would seem to be a paradoxical phenomenon. Nevertheless, under certain conditions such phenomena are found to occur. Before these conditions are defined generally, an instance of a psychological kind will be given to show how the tendency arises.

As part of a research on recruitment to nursing, a questionnaire was circulated to women employed in five kinds of work, viz.,

S. *Saleswomen*, chiefly shop assistants in departmental stores;
F. *Light Factory Workers*, employed on light assembly work, sorting and packing, cardboard box making, etc.;
T. *Teachers* in primary schools;
N. *Trained Hospital Nurses*, including sisters, charge nurses, staff nurses, etc.;
C. *Clerks*, chiefly shorthand typists, audit clerks, post office workers, etc., without specially responsible or administrative functions.

It will be convenient to use the term ' group ' of the persons, classified as above, and the term ' occupation ' in classifying their answers. The questionnaire itself concerned the comparative merits of the five occupations; and one section—the section to be discussed here—dealt more particularly with dress.

The instructions were: " On the whole, which of each pair is likely to be the better dressed off duty? Underline that one. If both seem about equal, underline both. If your knowledge is too limited to allow you to choose, do not underline either." Then follow the ten pairs obtained by taking the occupations two at a time, e.g., 1. School Teachers, Trained Hospital Nurses; 2. Trained Hospital Nurses, Clerks; etc.

In scoring each paired comparison, when one occupation is underlined and the other not, the former receives 2 marks, the latter none; when both or neither are underlined, each receives 1 mark. The marks given by each person to each occupation can be summed over the ten pairs; and each person thus provides five scores, one for each occupation. For instance, the scores assigned by four saleswomen are as shown in Table I.

TABLE I. SCORES ASSIGNED BY FOUR SALESWOMEN TO THE OCCUPATION SPECIFIED

Person Judging	S. Sales-women	F. Light Factory Workers	T. Teachers	N. Nurses	C. Clerks	Total
A	8	2	4	0	6	20
B	7	0	3	3	7	20
C	5	3	4	4	4	20
D	6	6	1	1	6	20

As each subject makes the same number of paired comparisons, the total of every row is the same (20). There are inconsistencies in a few responses, e.g., one person may place the clerks below teachers, teachers below nurses, and yet clerks above nurses when these are directly compared. In such a case each occupation contributes the same quantity towards the total score. In answering other questions (e.g., "Which group is likely to have had the better education"), a more objective judgment is possible ; and there is greater amount of agreement among the groups.

A first indication of the results is provided by summing the scores for each occupation within each group.

TABLE II. TOTAL SCORE ASSIGNED BY EACH GROUP TO THE SEVERAL OCCUPATIONS

Group	*Occupation* S	F	T	N	C	Total
S. Saleswomen	123	38	52	58	89	360
F. Light Factory Workers ..	123	42	47	45	103	360
T. Teachers	74	2	118	60	106	360
N. Nurses	78	26	86	95	75	360
C. Clerks	99	16	75	84	86	360
Total	497	124	378	342	459	1,800

As the row totals imply, 18 persons have been included in each group—a total of 90 in all groups combined. The column totals show that nurses and teachers are not generally thought to be as well dressed as saleswomen or clerks. But in the opinion of the teachers apparently teaching is the best-dressed occupation, and nursing in the opinion of the nurses. No one has a high opinion of factory work, but the highest is held by the factory workers. Intra-group loyalties are thus very much in evidence.

Still more interesting are the indications of similarity of opinion *between* the groups. Thus factory workers and saleswomen both show a high opinion of their own occupations and a low opinion of teaching and nursing. This might be interpreted as evidence of a feeling of inter-group loyalty. On the other hand for statistical purposes perhaps it would be better to say that there appears to be a certain affinity between these two groups. If degrees of this affinity could be measured in terms of generalized distances, phenomena of inter-group solidarity and class-consciousness might be brought within the scope of psychological measurement. From the standpoint of the original research into recruitment for nursing the question of chief interest is a specific one of this kind, namely, to compare the affinity of other groups with the Nurses.

As each occupation appears in four of the ten paired comparisons, the highest score that can be assigned to any occupation is 8. On theoretical grounds we may expect that the distribution of scores for any item will be similar to the binomial, and will approximate to the normal (except perhaps when the item has a high probability of being consistently preferred or disfavoured). When the scores given to any pair of occupations are plotted, the bivariate distribution also approximates to the normal. The same subject cannot give a score of 8 (or 0) to both occupations ; but this produces little disturbance.

Since a subject who esteems one occupation highly must rate the other occupations correspondingly low, and *vice versa*, the scores for the occupations tend to be negatively correlated. The variances and covariances of the scores, within and between groups, are shown in Table III. The correlations obtained by summing both parts of Table III are shown in Table IV. It will be seen that the only positive correlations are those between the occupations of Saleswomen and Light Factory

work and between Teaching and Nursing. The rest are negative, the largest being between Saleswomen and Teaching.

TABLE IIIA. TOTAL VARIANCES AND COVARIANCES WITHIN GROUPS (85 d.f.)

Group	S	F	T	N	C
S. Saleswomen	323·28	8·44	−194·78	− 99·17	− 37·78
F. Factory		227·78	− 98·44	−121·00	− 16·78
T. Teachers			354·11	16·72	− 77·61
N. Nurses				285·22	− 81·78
C. Clerks					213·94

TABLE IIIB. TOTAL VARIANCES AND COVARIANCES BETWEEN GROUPS (4 d.f.)

Group	S	F	T	N	C
S. Saleswomen	123·18	70·80	−140·62	−64·43	11·08
F. Factory		59·38	− 97·36	−25·20	− 7·62
T. Teachers			182·29	45·88	9·81
N. Nurses				93·18	−49·42
C. Clerks					36·16

TABLE IV. CORRELATIONS BETWEEN GROUPS

Group	S	F	T	N	C
S. Saleswomen	—	·2213	−·6854	−·3980	−·0799
F. Factory		—	−·4989	−·4435	−·0910
T. Teachers			—	·1389	−·1851
N. Nurses				—	−·4265
C. Clerks					—

II. DEFINITION OF NEGATIVE MATRICES

This negative tendency appears to originate in the instructions, which limit the persons in their choice of the occupations preferred. The sum of the ratings is fixed by the number of paired comparisons and is the same for all persons. In general, if the scores assigned to m objects are $x_1, x_2, \ldots x_m$, the instructions assign to each total, Σx, a constant value, say k. Now let Vx_1 denote the total variance of scores for the first object, Wx_1x_2 the total covariance between scores assigned to the first two objects, etc. Then $Vx_1 + Wx_1x_2 + \ldots Wx_1x_m = Wx_1 (x_1 + x_2 + \ldots x_m) = Wx_1k = 0$. Thus in Tables IIIA and IIIB the sum of each row and each column is 0. Moreover, since Vx_1 must be positive, $Wx_1x_2 + Wx_1x_3 + \ldots Wx_1x_m$ must be negative. Thus, a general tendency to negative inter-correlation arises whenever the condition $x_1 + x_2 + \ldots x_m = k$ is satisfied. In the psychological field and in sociometric studies the same tendency is to be expected wherever the method of paired comparisons or rank order is used.

With ranking, complications arise if the objects compared cannot be bracketed. Thus in an earlier experiment subjects were shown post-card sized reproductions of a set of paintings representing various styles, and were asked to arrange them in order of preference. Table V shows the frequency distribution for ranks assigned to two pictures—Constable and Reynolds. On the whole the subjects prefer the Constable ; but, if they like it, they tend to like the Reynolds too, and those who dislike one dislike both.

TABLE V. FREQUENCIES OF PAIRED RANKINGS

Rank for Reynolds	Rank for Constable							
	8	7	6	5	4	3	2	1
1	—	—	—	—	—	—	3	—
2	—	—	1	—	—	3	—	7
3	—	—	1	—	—	—	4	11
4	—	2	1	1	—	4	2	10
5	—	1	2	—	1	4	2	1
6	—	2	—	2	—	1	3	3
7	1	—	3	1	—	—	2	4
8	—	2	3	3	4	2	2	1

This association of preferences might have been more evident if subjects had been allowed to give an equal rank to both when they desired. As it is, a belt of zeros runs diagonally upward where cases are excluded. If x_c and x_r are used to denote the places assigned to Constable and Reynolds respectively, the general condition $x_1 + \ldots + x_c + \ldots + x_r + \ldots + x_m = k$ holds as before, but a series of supplementary conditions are introduced of the kind $x_1 \neq x_c \neq x_r \neq x_m$, etc. As a result, the bivariate distributions cease to resemble the normal form and become much more intractable statistically. If bracketing is permitted, such supplementary conditions are eliminated.

Sociometric data of the kind studied by Professor Moreno and his colleagues have a marked formal similarity to rankings or paired comparisons, and similar treatments may prove useful. It seems clear, however, that additional statistical problems are involved and the most that can be hoped for is that the treatment proposed here will go some way towards solving them. The present experiment may perhaps deserve to be called sociometric in a broader sense, as it is concerned with expressing differences between social groups metrically.

III. THE TRANSFORMATION PROPOSED

In the simplest case, where $m = 2$ and $x_1 + x_2 = k$, all points (x_1, x_2) lie on a single straight line, intersecting the axes at a distance of k from the origin. Variation along this line may be represented by a single measurement, $(x_1 - x_2)/\sqrt{2}$. The change of co-ordinates may be expressed in the usual way, $[x_1, x_2]\, T$, where $T = \begin{bmatrix} +1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix}$. When $m = 3$, the variations are confined to a single plane, and two measurements are sufficient to describe them. To effect the change any T may be used which has 3 rows and 2 columns, and the other characteristics defined below. Generally, where $x_1 + x_2 + \ldots x_m = k$, variation is confined to an $m - 1$ dimensional hyperplane of the m-dimensional space. T must have m rows and $m - 1$ columns ; and the axes it specifies must (i) be given by direction cosines, (ii) be orthogonal to one another, and (iii) be orthogonal to the dimension $(1, \ldots 1)$ in which no variation occurs. The algebraic equivalents of these geometrical properties are : (i) $t_{1i}^2 + \ldots t_{mi}^2 = 1$, (ii) $t_{1i}t_{1j} + \ldots t_{mi}t_{mj} = 0$, (iii) $t_{1i} + \ldots t_{mi} = 0$. A suitable T is the following :

$$T = \begin{bmatrix} +1/\sqrt{2} & +1/\sqrt{6} & \ldots & +1/\sqrt{m(m-1)} \\ -1/\sqrt{2} & +1/\sqrt{6} & \ldots & +1/\sqrt{m(m-1)} \\ 0 & -\sqrt{2/3} & \ldots & +1/\sqrt{m(m-1)} \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & -\sqrt{(m-1)/m} \end{bmatrix}$$

Although the transformation matrices given in Burt's *Factors of the Mind* (pp. 303–6) are similar in some respects, they do not have the essential properties and cannot be used here. The process $X\,T$ transforms a score matrix X, with m scores for each subject, into a matrix with $m - 1$ scores. Each new score will be a weighted difference between one score and the sum of the scores considered previously. Sums of scores and averages will be similarly transformed.

When the appropriate T is evaluated and applied to the data in Tables II and III, they are converted into Tables VI and VII. No importance is attached here to the order in which the successive comparisons ($S - F$, $S + F - T$, etc.) are taken. In other experiments one order may perhaps be more closely related to the problem at issue than another. The decision involved may perhaps be compared with that which has to be made in an analysis of variance, where the order in which the degrees of freedom are broken down is decided by the hypotheses under consideration.

TABLE VI. SUMS OF MEASUREMENTS OF DIFFERENCES BETWEEN SCORES OF OCCUPATIONS

Group	1	2	3	4
Sales	60·10	23·26	11·27	— 19·01
Factory	57·28	28·98	22·23	— 34·66
Teachers	50·91	−65·32	4·05	— 38·01
Nurses..	36·77	−27·77	−27·42	— 3·35
Clerks	58·69	−14·29	−17·89	— 15·65
Total	263·75	−55·14	— 7·76	−110·68

TABLE VII. VARIANCES AND COVARIANCES OF THE MEASUREMENTS

A. Within Groups

	1	2	3	4
1	267·09	83·18	— 13·54	16·60
2		526·17	107·68	— 45·95
3			343·60	— 36·52
4				267·41

B. Between Groups

	1	2	3	4
1	20·48	43·40	28·22	−14·78
2		334·17	87·41	7·38
3			94·30	−52·13
4				45·19

IV. TREATMENT OF THE DATA

The transformation changes the dimensions in which the multivariate dispersion is recorded, and renders the results amenable to treatment by established methods. The analysis follows recognized lines. The proportion of the variance within groups to the total

is given by $\Lambda = \frac{\text{determinant of dispersion within groups (Table VIIA)}}{\text{determinant of the total dispersion (Tables VIIA and VIIB)}} = \cdot 41124$. Hence, $\chi^2 = 75 \cdot 09$ with 16 d.f. Thus the groups evidently vary significantly. The generalized distance between each pair of groups may be examined next. The discriminant functions so found are shown in Table VIII :

TABLE VIII. DISCRIMINANT FUNCTIONS

Pair	Discriminant	Significance
1. Saleswomen—Light Factory Workers	·3037	$P > \cdot 20$
2. Saleswomen—Teachers	4·7464	$P < \cdot 01$
3. Saleswomen—Nurses	2·3938	$P < \cdot 01$
4. Saleswomen—Clerks	1·0896	$\cdot 10 > P > \cdot 05$
5. Light Factory Workers—Teachers	4·6644	$P < \cdot 01$
6. Light Factory Workers—Nurses	3·6921	$P < \cdot 01$
7. Light Factory Workers—Clerks	1·9097	$P < \cdot 01$
8. Teachers—Nurses	3·9051	$P < \cdot 01$
9. Teachers—Clerks	2·7666	$P < \cdot 01$
10. Nurses—Clerks	·7169	$P = \cdot 20$

Some stratification now appears between the groups. When arranged in the order (1) Light Factory Workers, (2) Saleswomen, (3) Clerks, (4) Nurses, (5) Teachers, each group is placed closest to those from which it is least well discriminated. But this rough arrangement does not show all the differences : e.g., the Teachers form a group unmistakably distinct from the rest, yet are less remote from the Clerks than from the Nurses.

To make an adequate graphical representation of relations we need to solve the usual type of determinantal equation, namely :—

$$\begin{vmatrix} 5\cdot5296-\mu & -3\cdot6618 & 8\cdot4429 & -8\cdot3248 \\ 4\cdot7374 & 54\cdot4970-\mu & 7\cdot5718 & 6\cdot5616 \\ 5\cdot3415 & 5\cdot7529 & 19\cdot8977-\mu & -13\cdot7791 \\ -3\cdot4987 & 12\cdot7726 & -13\cdot0765 & 14\cdot1272-\mu \end{vmatrix}$$

$= \mu^4 - 94\cdot0515\mu^3 + 2{,}260\cdot7486\mu^2 - 7{,}492\cdot4744\mu + 4{,}053\cdot6008 = 0$. The four roots are ($\mu_1$) 56·1292, ($\mu_2$) 34·1086, ($\mu_3$) 3·1393, ($\mu_4$) ·6745. They define the principal axes along which variation occurs between the groups. Applying the χ^2 test (with 7, 5, 3, and 1 degrees of freedom) it appears that the first two dimensions are sufficient to represent the variation between the groups ; the other two are negligible. The positions of the 5 mean points in the two-dimensional space are plotted in Fig. 1. For this purpose we must first find the appropriate *k*-coefficients, and then calculate the means (Table IX).

TABLE IX. *k*-COEFFICIENTS AND MEANS

Variable	A. *k*-coefficients μ_1	μ_2	Variable	B. Means μ_1	μ_2
1. }			1. S	1·806	−7·005
2. } S−F	−0·362	−1·263	2. F	2·070	−9·658
3. S+F−T	3·055	·473	3. T	−14·061	−5·776
4. S+F+T−N	·078	−2·161	4. N	− 5·744	−0·379
5. S+F+T+N−C	·931	1·937	5. C	− 4·493	−4·030

V. INTERPRETATION OF RESULTS

Only a summary of the main results and a few brief comments will be given here. Let us take the differences between the Nurses and the other ' groups ' first. As will be seen from Table VIII, Clerks are nearer to them than any other group—so near, indeed, that the difference is non-significant. Saleswomen come next. Light Factory Workers are further away still. Teachers form an isolated group, widely separate from the others. On this evidence Clerks are more likely to be attracted to nursing than women in any other occupation. The differences between ' occupations ' may be inferred from Table II. The general opinion is that Saleswomen are most likely to be well dressed, and Light Factory Workers least. Yet, as Table VIII shows, Saleswomen and Light Factory Workers, though widely dissimilar as occupations, nevertheless resemble each other closely in group-opinion.

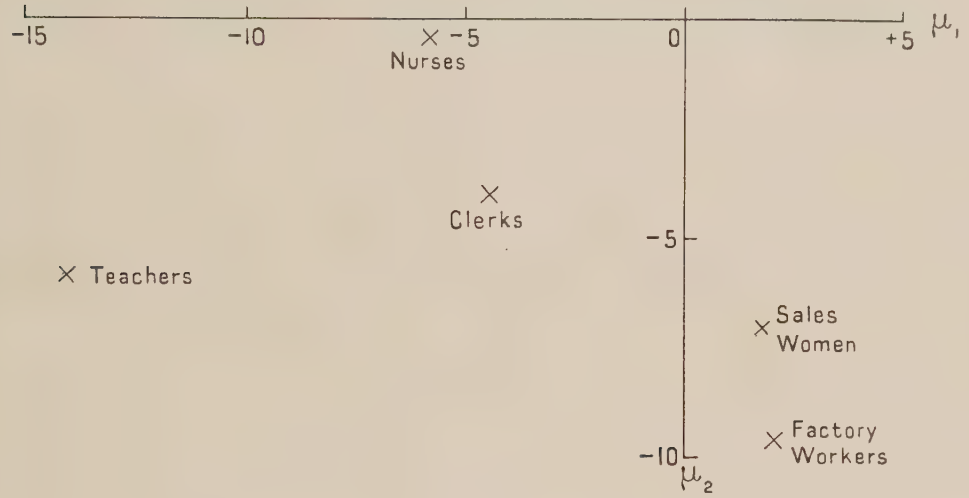


FIG. 1. The dispersion of the groups in the two main dimensions.

NOTE.—*As the minor dimensions are omitted from the diagram, which is based on the last two columns of Table IX only, the distances here shown do not tally exactly with those indicated by the figures in Table VIII (column headed ' Discriminant ').*

There is evidently no simple relation between the observed dispersion among the groups and the characteristics of the occupations to which the questionnaire refers. If the data had referred to length of education or degree of responsibility of the work, different proportions of the total dispersion might have been found, yet constellations resembling Fig. 1 might still have emerged. What is decided by the choice of topic and the system of scoring is the *space* within which the variation between and within the groups appears. Within that space the responses given by members of each group determine where their own mean point occurs ; and the *points* are therefore decided by the selection not of the topic but of the subjects.

To define the dimensions μ_1 and μ_2 in psychological terms, two methods of investigation might seem practicable : first, to examine the relation between these dimensions and the variables originally recorded ; from this it should be possible to describe the major axes along which variation occurs : secondly, to compare information about the groups from other sources with their position along the two dimensions to be defined. However, it seems over-ambitious to attempt any precise interpretation of the present data along these lines. A repetition of the analysis, with data from other parts of the questionnaire, might not confirm the stability of the configuration shown ; and the addition of new groups might change the direction of the axes.

VI. ALTERNATIVE TREATMENTS

Any method of analysis which treats the original scores as independent may cause difficulties. Some of these (e.g., the determination of degrees of freedom) may not prove intractable ; others may prove more troublesome (e.g., as a result of the limitation to $m - 1$ dimensions it is to be expected that the determinants of the matrices in Table III will have a zero latent root). Whether transformation is necessary before other treatments are applied may be open to doubt. The data in Tables II and III would seem suitable for an analysis of variance as they stand.

As we have seen, group loyalty or self-esteem is a quality exhibited to some extent by all the groups ; it may therefore appear in each of the dimensions, and so cannot be associated with any particular μ. To show how far it is sufficient to account for the differences observed, or whether the groups vary in the amount they exhibit, an analysis of variance may be employed.[1]

There are 450 scores altogether. Their sum is 1,800, the sum of their squares 10,042. Their total variance is thus 2,842. As the sums for individuals and for groups do not vary, it has only 360 d.f. These are divisible into variance between the occupations, group-and-occupation interactions, and residual variance (person-and-occupation interactions). The group-and-occupation interactions attributable to group loyalty may be isolated by extracting the scores in the leading diagonal of Table II from the remainder, as shown below :

TABLE X. SCORES AND MEANS

	Occupation					Total
	S	F	T	N	C	
(A). *Scores* assigned by						
(i) those in the occupation ..	123	42	118	95	86	464
(ii) others	374	82	260	247	373	1,336
Total	497	124	378	342	459	1,800
(B). *Means* assigned by						
(i) those in the occupation ..	6·8	2·3	6·6	5·3	4·8	5·2
(ii) others	5·2	1·1	3·6	3·4	5·2	3·7
General Mean ..	5·5	1·4	4·2	3·8	5·1	4·0

The general effect of self-esteem may be inferred from the totals, and variations in its amount from the figures for separate occupations. The remaining group-and-occupation interaction can be obtained from the other data in Table II. The resulting analysis is as follows (Table XI) :

TABLE XI. ANALYSIS OF VARIANCE

Source of Variance	Sum of Squares	d.f.	Mean Square
Between occupations	943·5	4	235·8
Between groups :			
(*a*) General effect of self-esteem	150·2	1	150·2
(*b*) Differences in amount in different groups.. ..	85·3	3	28·4
(*c*) Remaining interactions	258·7	12	21·5
Within groups (error)	1,404·3	340	4·1
Total	2,842·0	360	

[1] I owe this suggestion to Professor M. S. Bartlett, who, in the course of correspondence about my original draft, also very kindly submitted in outline a worked example.

In applying the usual test of significance, it has to be assumed that the variances within the occupations are equal to one another, and independent. This is equivalent to assuming that the matrix in Table IIIA does not differ significantly from a 5×5 matrix with each term in its leading diagonal = 280·86 and all the remaining terms = −70·216. A test proposed by M. S. Bartlett can be employed to verify this hypothesis ; but as it involves the use of determinants, the transformation is required. The determinant of Table VIIA, say Δ_1, is compared with that of the transformed matrix, say Δ_2, which has 351·08 for each term in its leading diagonal, and zero elsewhere. We take $\chi^2 = -\left\{ n - \frac{1}{6}\left(2p + 1 + \frac{p}{2}\right)\right\} \log_e \frac{\Delta_1}{\Delta_2} = 26{\cdot}82$. This has $p - 1 + \frac{1}{2}p\,(p-1)$ d.f. (since Δ_2 is deduced by hypothesis). Here n (d.f. of error variance in each dimension) = 85, and p (the number of dimensions in which variation occurs) = 4. The d.f. therefore = 9, and P is < ·01. The hypothesis has consequently to be rejected ; and doubts are therefore thrown on the inferences drawn from Table X. However, they can hardly be ignored, since the variance ratios are beyond the region of doubtful significance.

The main conclusions suggested are as follows. Group loyalty contributes largely to the dispersion between the groups ; and the amount varies from group to group. The remaining group and occupation interactions are also large. This and the doubtful test of significance provide a justification for the analysis made in Section IV ; and, indeed, logically the analysis of variance should perhaps have preceded the analysis of dispersion.

Other methods of analysis seem less likely to show the relevance of the data to the initial problem of recruitment. Burt's method of correlating persons is too cumbersome ; factor analysis appears unlikely to suggest useful conclusions, and raises a number of questions. In an ordinary correlation table absence of association between two variables is indicated by a correlation of zero ; in the negative matrix, however, absence of association is indicated by a correlation of $-1/(m-1)$. This leads to results which become more and more perplexing as more and more factors are extracted. And tests of significance appear impossible. Transformation eliminates all these difficulties.

If a Thurstone-type analysis is applied mechanically to a negative matrix, the communalities adopted must be negative, and the first factor extracted is an imaginary one. All its loadings are coefficients of $\sqrt{-1} = i$. With Table IV we should obtain the saturations shown in Table XII.

TABLE XII. RESULTS OF FACTOR ANALYSIS

Occupation	S	F	T	N	C
Factor loadings	·465i	·386i	·678i	·595i	·369i

The introduction of an imaginary general factor may perhaps be open to no more serious objections than many other methods which treat residual variances as specific factors. A plausible psychological interpretation is not lacking for the imaginary factor. The test situations in which it appears are characterized by a typical form of mental activity, unlike the processes of educing relations associated with positive g. Here thinking is disjunctive, and thus negates the relation-finding process. We may perhaps regard the type of thought processes involved as dialectical (in the Hegelian and Marxist sense). But the introduction of imaginary quantities seems hardly defensible statistically, since the observation is confined to a real field : whether it does in point of fact yield anything of value, the reader may be left to decide.

In conclusion I should add that this paper owes much to Miss M. Farnworth, who collected the data, to M. Ash for mathematical assistance, to Professor M. S. Bartlett for suggestions, particularly on the analysis of variance, and to Sir Cyril Burt for other helpful comments and criticisms.

VII. SUMMARY

1. When m objects are compared repeatedly, as in experiments involving ranking or paired comparison, a score may be allotted to each object to indicate the position assigned to it relative to the others. Except in extreme cases, provided certain conditions are observed, the univariate and bivariate distributions tend to be normal.

2. The bivariate distributions, however, tend to yield negative correlations. The reasons for this tendency have been explained. Difficulties that arise in applying established statistical procedures to such matrices can be avoided by using a transformation matrix, which is described.

3. The data illustrating this procedure are taken from an enquiry into recruitment for nursing. The occupations of saleswoman, light factory work, teaching, nursing, and clerical work have been compared in regard to dress by 90 women —18 in each group. Statistical analyses applied after transformation yield measures of the divergence of opinion between the persons in the different occupations, and show in which occupation the opinions most closely resemble those of the nurses.

4. It is suggested that the technique described may indicate fresh uses for the methods of paired comparison and rank order, especially in sociometric studies of inter-group solidarity and class consciousness.

A REPLY TO MR. SLATER'S CRITICISMS

Mr. Slater has compared his transformation matrix with the triangular transformation matrices which I have described in *Factors of the Mind* and earlier papers. He concludes (p. 13) that my matrices " do not have the essential properties and cannot be used here." Since we have not been able to reach complete agreement either on this point or on the value of a factorial approach, it has been suggested that a supplementary note might be helpful.

His matrices and mine are special forms of what I called *equilibrated* orthogonal matrices, i.e., their rows and/or columns sum to zero. The triangular transformation to which he alludes was merely a standard numerical version of that described by Yule (*Theory of Statistics*, 1912, p. 331). The formula which I gave for constructing the complete (square) matrix (e.g., this *Journal*, I, p. 8) seems *identical* with Mr. Slater's. Its chief use was to reduce a singular group factor matrix of order $n + 1$ to a bipolar factor matrix whose order was equal to its rank. Without such a transformation the determinant of the factor matrix would be zero, and no inverse could be calculated for deducing factor measurements. That seems to be the *real* reason here : for, without such a transformation, the determinants in Sections IV and VI would vanish. Much the same type of transformation has been used in theoretical deductions by Burnside, Irwin, and David, and doubtless other mathematical writers, for much the same purpose.

In the passage which Mr. Slater quotes my object was to convert a bipolar matrix to a group factor matrix, by getting rid of the negative *saturations*. Mr. Slater says his object is to get rid of the negative *covariances*. Now the tendency to negative covariances in tables such as he has described may arise from *two* causes. (1) The restriction in the instructions (which he regards as the essential cause, cf. p. 11) would in general produce only an ordinary *bipolar* matrix : its effect is equivalent to removing the general factor. Bipolar tables of this kind occur, not only with paired comparisons or with ranking, but with *any* method of marking which produces a ' doubly centred ' measurement matrix, e.g., with ratings standardized in deviation form (cf. *Factors of the Mind*, pp. 380–1). (2) But a ' *negative* covariance matrix ' (as I understand the phrase) denotes something more than a *bipolar* matrix ; it implies a matrix in which *all* the covariances are negative. Such a matrix is commonly the result of specific self-relation factors. Of these perhaps the most frequent types, in psychological research, are the factors of self-approval or self-preference. It is this that really causes most of Mr. Slater's difficulties. In my view, therefore, if the ultimate aim was to assess the objective ' affinity ' or ' distance ' between the groups, then these factors should have been avoided or removed, just as we try to avoid or remove specific factors from the self-correlations when we wish to isolate the common or group factors in an ordinary correlation table.[1]

[1] This criticism, as well as that in paragraph ii, has now been largely met by the section on ' Alternative Treatments ' (pp. 16-17).

Their influence is most readily studied if we take extreme or limiting cases. (i) Consider first the hypothetical case in which there is no correlation, positive or negative, between the ratings, except what is induced by self-approval. Without standardization the correlation matrix would then be an $n \times n$ unit matrix with rank n. On factorizing this we should obtain a triangular orthogonal matrix such as I have described (cf. this *Journal*, II, p. 112, Table V). If the columns are numbered backwards, the first or general factor will have positive saturations of $1/\sqrt{m} = 1/\sqrt{n}$. The remaining columns will have m negative elements, each $= -1/\sqrt{m(m+1)}$, and a diagonal element $= +m/\sqrt{m(m+1)}$. They thus comprise a set of 'equally balanced' bipolar factors (this *Journal*, I, p. 8). If, however, the ratings are standardized in deviation form, the general factor will be thereby suppressed. The covariance matrix will consequently be negative, each covariance having the value $-1/n$; and the factor matrix will be identical with Mr. Slater's transformation matrix. An empirical matrix of the type with which Mr. Slater is concerned will doubtless have small correlations or covariances of its own in addition to those so produced. Pre- and post-multiplying a covariance matrix of this kind by such a transformation matrix will therefore be equivalent to removing the negative covariances due to the influence of self-approval. But the factor itself will not thereby be removed.

(ii) Consider now the case in which the correlations would have been perfect, except for the influence of the factors of self-approval. We may first suppose that such a factor merely adds some positive amount to the mark that would otherwise be given, and that the marks are not further standardized, so that the *absolute* amounts of the unaffected marks are left the same for all. The best way to deal with this problem would be by an analysis of variance, in which the effects of self-approval could be treated as a single factor. If, however (following Mr. Slater's procedure), we covariate traits (occupations), then we automatically remove the similarity between persons. Hence the covariance matrix will again be negative, with the same values as before ; and the transformation will have the same result.

(iii) Finally, suppose that the effect of self-approval is to raise the judge's own occupation to the first place in all cases, but that the marks are now standardized, so that the *relative* amounts are the same for all. Of the three simple causal schemes this would seem to be the one that most nearly represents the situation Mr. Slater has in mind. In such a case there can, I think, be no question whatever that we should correlate persons (or groups of persons), not traits or items (occupations). Mr. Slater objects that my proposal to correlate persons would be 'too cumbersome'; but the calculations entailed are far shorter than those he himself has used. In the simplified case we should

TABLE XIII. CORRELATIONS BETWEEN PERSONS OR GROUPS

Persons	A. Correlations					B. Triangular Factor Matrix				C. General and Imaginary Factors	
	A	B	C	D	E	I	II	III	IV	I	II
A	(1·0)	·9	·7	·4	·0	1·0	·0	·000	·000	1·0	·0*i*
B	·9	(·8)	·6	·3	− ·1	·9	− ·1	−·173	−·387	·9	·1*i*
C	·7	·6	(·4)	·1	− ·3	·7	− ·3	−·520	·387	·7	·3*i*
D	·4	·3	·1	(−·2)	− ·6	·4	− ·6	·693	·000	·4	·6*i*
E	·0	−·1	−·3	−·6	(−1·0)	·0	1·0	·000	·000	·0	1·0*i*

TABLE XIV. FACTOR ANALYSIS OF COVARIANCES BETWEEN OCCUPATIONS

Occupations	Factors			
	IA	IIA	IB	IIB
Saleswomen	15·5	7·6	14·6	5·4
Factory Workers	10·6	4·0	8·5	− 1·3
Teachers	−18·6	−10·7	−18·6	− 8·1
Nurses	−13·7	10·7	−13·2	15·4
Clerks	6·2	−11·7	3·3	− 6·7

then get an ' isoclinal ' table like Table XIIIA. And once again the covariances or correlations can be expressed by a simple algebraic formula. If the persons are numbered in order of agreement with the average marks, then the correlation of the pth and qth persons will be $r_{pq} = 1 - \frac{6\{p(p-1) + q(q-1)\}}{n(n^2-1)} = 1 - a_p - a_q$ (say) $= (1 - a_p)(1 - a_q) - a_p a_q$. This is an ' isoclinal ' not a ' synclinal ' hierarchy. If we retain self-correlations of 1·00, and factorize the table by the triangular method, we should obtain (with 5 variables) the saturations printed in Table XIIIB. If, however, we substitute reduced self-correlations (shown in brackets in Table XIIIA), we should obtain the saturations printed in Table XIIIC. In fact, as the above equation demonstrates, we can always express such a table as the result of two factors, one of which will be a general factor and the other an imaginary factor, the numerical coefficients always summing to 1·00 for the same variable.

It was partly to meet such cases that I suggested [1] that the introduction of complex numbers might prove to be as useful in the psychological factor analysis as it has been in quantum mechanics. Mr. Slater proposes a psychological interpretation : he argues that, while a factor with ' real ' elements represents integrative or ' relational ' thinking, a factor with ' imaginary ' elements might represent dissociative or ' disjunctive ' thinking. I would suggest a mathematical interpretation : here as elsewhere we may regard an imaginary number as representing a measurement along a dimension at right angles to that indicated by the corresponding real number.

If we *must* covariate items (as in Table III) rather than persons or groups, then the most plausible way to factorize the results would be to use the ordinary method of simple summation, with the signs of N and C reversed, as usual, for purposes of computation : (I cannot agree that, if Mr. Slater's ' negative ' matrix is ' factorized mechanically,' the " communalities *must* be negative "). The results are shown in Table XIV. To allow for the specific factors of self-approval we can use ' reduced ' variances : the loadings then obtained are shown in the last two columns. These results may be compared with Mr. Slater's figures in Table IXB. The graph derived from my figures would be very similar to his, and the figures themselves are far easier to compute. The chief difference is that, as a result of the change in the first factor (cf. Table IXB and Table XIV, col. IA), the Clerks are now classed with the industrial or commercial workers, rather than with the semi-professional groups like the Teachers or Nurses. That after all is what we should expect from ordinary observation of their mode of dressing : but otherwise the changes are quite slight. C. Burt.

[1] Cf. ' Correlations between Persons,' *Brit. J. Psych.*, XXVIII, 1937, p. 69 : ' The Unit Hierarchy and its Properties,' *Psychometrika*, III, 1938, p. 156, and refs.

FACTOR ROTATION BY THE METHOD OF EXTENDED VECTORS[1]

By JOHN SUTHERLAND
Moray House, University of Edinburgh

I. *Centroid Factor Analysis.* II. *Orthogonal Rotation of the Centroid Factor Loadings.* III. *Rotation by the Method of Extended Vectors.* IV. *Second Order Factors.* V. *Conclusions.* VI. *Summary.*

I. CENTROID FACTOR ANALYSIS

The object of the following article is to describe a factorial analysis into oblique factors, and subsequently into second order orthogonal factors, and to evaluate the possibilities of the method. No attempt will be made to describe either the tests or the conditions of the experiment beyond what is necessary for an understanding of the factorial analysis. During an investigation into the arithmetical and other abilities of children a battery of 18 tests was applied to all the 9-year-old children who had reached or passed the fifth year in four city primary schools. The results of these tests together with three pieces of information supplied by the schools yielded 21 variables. In this present article the results for 136 boys only will be considered.

The correlations are shown in Table I (overleaf). This matrix was factorized by the centroid method using the highest correlation in each column as communality and without iteration. Five factors accounted for 80 per cent. of the variance. Coombs' criterion (1) and McNemar's (2) showed that the residuals were still significant ; but, as later factors continued to extract less than 2 per cent. of the variance, only the first five are considered here.[2] They are shown in Table II.

TABLE II. MATRIX OF CENTROID LOADINGS IN FIVE FACTORS F_0

No.	Test	I	II	III	IV	V	h^2
1	Arithmetic Quotient	·793	—·314	—·249	—·084	—·027	·798
2	I.Q. (Picture Test)	·744	·335	—·100	—·187	—·171	·741
3	Teachers' Arithmetic Estimate ..	·842	—·278	—·094	·080	—·058	·804
4	Mechanical Arithmetic	·800	—·443	—·105	—·099	—·062	·861
5	Problem Arithmetic	·884	—·173	·245	·203	—·106	·923
6	Mechanical Arithmetic	·811	—·458	·071	—·308	·053	·970
7	Problem Arithmetic	·882	—·173	·183	·076	·021	·847
8	Speed Mechanical Arithmetic ..	·746	—·552	—·103	—·064	·115	·890
9	Speed Problem Arithmetic ..	·811	—·191	·138	·162	·165	·767
10	Spatial Problem Arithmetic ..	·836	·101	·213	·145	·116	·788
11	Verbal Intelligence	·920	·072	·104	·017	·050	·865
12	Non-Verbal Intelligence	·818	·221	—·077	·170	·141	·774
13	Jenkins' Test (A)	·813	·279	—·263	·202	·224	·899
14	Matrix Test	·682	·353	—·088	—·045	—·214	·646
15	Jenkins' Test (B)	·781	·348	—·211	·319	·091	·886
16	Space Test (A)	·631	·233	—·131	—·156	—·112	·507
17	Space Test (B)	·663	·279	·081	·213	—·147	·592
18	Number Series	·830	·076	·226	·112	—·259	·826
19	Letter Series	·814	·132	·108	·141	—·284	·794
20	Vocabulary..	·812	·089	·170	—·344	·271	·887
21	Interpretation	·782	·069	·194	—·231	·117	·722
	Percentage of Variance ..	63·7	7·9	2·7	3·3	2·4	80·0

[1] Based on extracts from a thesis submitted for the degree of Ph.D. in the University of Edinburgh.
[2] The matrix was also analysed by Lawley's Method of Maximum Likelihood (3) using the centroid loadings as a first estimate. Lawley's own test of significance for the residuals left after the extraction of five factors also indicated that the residuals were significant.

TABLE I. INTER-CORRELATIONS OF SCORES OBTAINED BY 136 BOYS ON 21 TESTS

Name of Test	No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Arith. Quotient	1		·602	·742	·790	·686	·764	·696	·842	·648	·527	·711	·570	·612	·415	·563	·412	·467	·597	·600	·577	·591
I.Q. (Pic. Test)	2			·533	·467	·576	·448	·554	·377	·459	·607	·743	·692	·620	·619	·595	·606	·627	·630	·663	·640	·623
Teachers' Arith. Est. ..	3				·800	·794	·798	·777	·788	·726	·643	·748	·633	·662	·479	·595	·521	·461	·654	·697	·608	·619
Mechanical Arith. ..	4					·733	·904	·754	·843	·684	·568	·676	·557	·506	·443	·482	·404	·433	·635	·598	·592	·599
Problem Arith.	5						·759	·860	·686	·815	·798	·834	·712	·632	·539	·663	·545	·375	·780	·727	·670	·727
Mechanical Arith. ..	6							·797	·867	·672	·611	·714	·542	·547	·442	·273	·418	·505	·624	·563	·769	·636
Problem Arith.	7								·740	·739	·797	·810	·700	·588	·515	·645	·500	·507	·788	·710	·683	·714
Speed Mech. Arith. ..	8									·739	·533	·658	·480	·455	·322	·437	·309	·382	·531	·501	·586	·554
Speed Prob. Arith. ..	9										·776	·747	·615	·633	·410	·613	·421	·487	·666	·617	·643	·631
Spatial Prob. Arith. ..	10											·776	·702	·686	·576	·678	·555	·598	·739	·691	·669	·636
Verbal Intell.	11												·772	·757	·595	·738	·539	·621	·783	·772	·780	·763
Non-Verbal Intell. ..	12													·861	·596	·723	·460	·604	·648	·679	·630	·631
Jenkins' Test (A)	13														·660	·900	·574	·563	·588	·618	·629	·591
Matrix Test	14															·660	·614	·492	·632	·672	·537	·510
Jenkins' Test (B)	15																·567	·563	·636	·665	·592	·561
Space Test (A)	16																	·533	·489	·435	·537	·487
Space Test (B)	17																		·547	·538	·610	·543
Number Series	18																			·844	·583	·627
Letter Series	19																				·573	·595
Vocabulary	20																					·834
Interpretation	21																					

II. ORTHOGONAL ROTATION OF THE CENTROID FACTOR LOADINGS

These factors were then rotated orthogonally in pairs. The guiding principle in all cases was rotation of the axes to positions which could be interpreted psychologically. Simple structure as such was not directly sought ; but, wherever possible, axes were placed through a cluster or line of test points. The results are shown in Table III.

TABLE III. ORTHOGONALLY ROTATED CENTROID FACTOR LOADINGS

No.	Test	I General	II Number	III Space	IV Reasoning	V Verbal
1	Arithmetic Quotient	·466	·531	·397	·151	·342
2	I.Q. (Picture Test)	·772	—·040	·313	·172	·119
3	Teachers' Arithmetic Estimate ..	·492	·527	·264	·342	·317
4	Mechanical Arithmetic	·480	·656	·246	·135	·350
5	Problem Arithmetic	·597	·455	—·041	·527	·283
6	Mechanical Arithmetic	·565	·623	·078	—·030	·506
7	Problem Arithmetic	·588	·412	·018	·406	·409
8	Speed Mechanical Arithmetic ..	·324	·689	·219	·309	·496
9	Speed Problem Arithmetic ..	·432	·371	·042	·446	·493
10	Spatial Problem Arithmetic ..	·587	·124	·008	·501	·421
11	Verbal Intelligence	·693	·191	·130	·402	·413
12	Non-Verbal Intelligence.. ..	·530	·013	·296	·503	·387
13	Jenkins' Test (A)	·460	—·058	·480	·518	·429
14	Matrix Test	·693	—·055	·288	·280	·031
15	Jenkins' Test (B)	·472	—·087	·430	·629	·274
16	Space Test (A)	·614	·005	·304	·134	·133
17	Space Test (B)	·560	—·004	·112	·510	·068
18	Number Series	·731	·254	—·009	·467	·098
19	Letter Series	·709	·210	·107	·482	·047
20	Vocabulary	·701	·059	·042	·047	·623
21	Interpretation	·690	·118	·010	·136	·463

Factor I, with loadings ranging from ·324 to ·772, may be interpreted as a *general factor*. Factor II, with high loadings in all the arithmetic tests and low or almost zero loadings in the others, appears to be a *number factor*. Factor III has highest loadings in all the tests requiring spatial imagery : it may be called a *space factor*. Factor IV has high loadings in problem arithmetic, non-verbal and verbal intelligence, Jenkins' test, the three-dimensional space tests, and the two series tests ; it may be called a *reasoning factor*. Factor V has highest loadings in those tests requiring knowledge of the English language ; it may be called a *verbal factor*.

It must be remembered that these are not the only possible positions for the axes. They are simply the positions which, in the subjective opinion of the author, appear to give the most satisfactory psychological interpretation.

III. ROTATION BY THE METHOD OF EXTENDED VECTORS

Following the procedure described by Thurstone, the extended factor matrix E_c was formed, and post-multiplied by a Landahl matrix M_0 to give the matrix E_0, shown in Table IV (4) (5).[1] The pairs of columns of this matrix E_0 were plotted on graph paper and new positions for the axes were sought. As there were five columns of entries in E_0 (apart from column one) there were ten possible pairs.

[1] For reasons of space it has not been possible to reproduce in full the computations involved in this part of the article. They are available in typescript form for anyone who wishes to see them.

TABLE IV. EXTENDED FACTOR LOADINGS ($E_0 = E_c M_0$)

	I	II	III	IV	V	VI
1	1	·093	·264	·540	·646	·694
2	1	·849	·230	·180	·325	·650
3	1	·152	·424	·631	·466	·564
4	1	−·048	·458	·508	·604	·714
5	1	·272	·731	·598	·233	·403
6	1	−·058	·650	·238	·749	·657
7	1	·272	·671	·501	·413	·379
8	1	−·215	·493	·582	·797	·579
9	1	·236	·647	·614	·513	·226
10	1	·555	·641	·487	·374	·177
11	1	·517	·527	·412	·428	·352
12	1	·688	·305	·583	·450	·207
13	1	·754	·090	·666	·557	·167
14	1	·909	·219	·315	·173	·617
15	1	·845	·114	·758	·341	·177
16	1	·777	·184	·223	·400	·651
17	1	·823	·458	·579	·030	·343
18	1	·529	·662	·458	·072	·513
19	1	·592	·526	·513	·055	·548
20	1	·545	·603	·016	·772	·300
21	1	·526	·642	·115	·582	·370

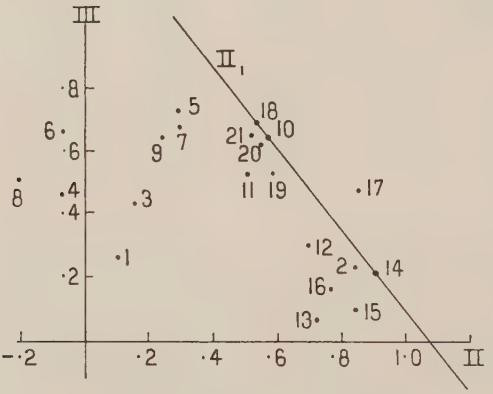


FIG. 1. Extended factor loadings.

On the graphs linear arrangements of the points were sought. The difference between this method and the orthogonal rotation by pairs is that the new lines may be placed through any linear arrangement of points on the diagram: (see Fig. 1). They do not have to pass through the origin. Five lines have to be chosen to give the new positions of the five axes. The lines were carefully chosen so that as far as possible the following conditions were fulfilled:

(1) They passed through or near to as many points as possible;

(2) They bounded the configuration of points, i.e., all the points which the line did not pass through lay on one side of it. This ensured that all the loadings would be of the same sign;

(3) The new projections produced were capable of psychological interpretation.

The coefficients of the terms in the equations of these lines, when multiplied by suitable constants to convert them from direction numbers into direction cosines, form the rotating matrix. Multiplication by this matrix carries the extended factorial matrix E_0 into the next matrix E_1.

Thurstone then repeats the whole process. On plotting the columns of the matrix E_1 in pairs, he usually finds that many of the points lie on or near one of the axes. He still considers it possible, however, to improve the position of the axes until all, or almost all, the points are clustered along the axes. In the example given in *Multiple Factor Analysis* this takes four rotations and produces simple structure. In the present example it was found that, by carefully choosing the position of the axes in the first rotation, one rotation was sufficient to carry the reference vectors to the best possible position.

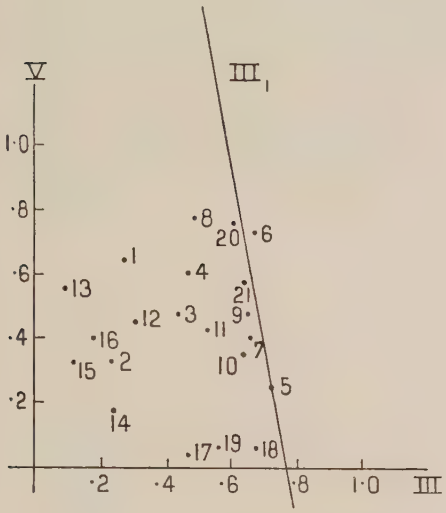


FIG. 2. Determination of vector III_1.

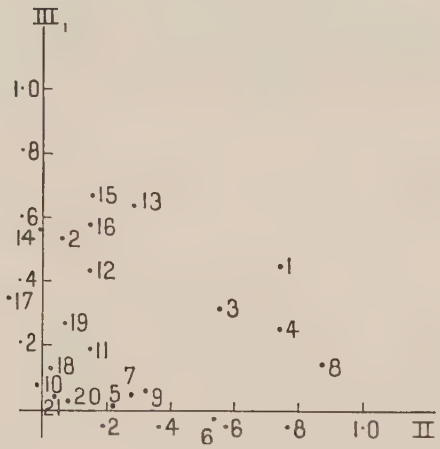


FIG. 3. Points referred to vectors II_1 and III_1.

Consider for example axis II. Fig. 1 shows the twenty-one test points plotted with reference to axes II and III. We now want a position for one of the reference vectors in which it will pass through as many points as possible, will bound the configuration of points, and yield a matrix of projections which it is possible to interpret psychologically.

With this in view, vector II_1 was placed in the position shown in Fig. 1. A position for each of the other vectors was similarly found. The diagram from which the position of the vector III_1 was obtained is shown in Fig. 2. If now we plot the test projections on II_1 and III_1, we obtain the diagram shown in Fig. 3. In this figure it is seen that the test points 14, 2, 19, 18, 10, 20, 21 now all lie on or near the axis III_1, that is, have zero projections on axis II_1 and test points 21, 20, 5, 7, 9, 6 all lie on or near the axis II_1, that is, have zero projections on III_1.

Now if we wish to continue to satisfy the conditions which determined the position of axes II_1 and III_1 in the first rotation, we cannot alter their position in this new diagram. It might be suggested, for instance, that we should place axis II_2 through test point 17 ; but we have already decided to put up with a small negative projection for test 17 in order to give as many zero projections as possible on axis II_1.

Thurstone's example is artificially constructed, with 5 factors and 25 tests. He starts with a simple structure having 14 zero loadings in each factor, rotates these factors to a new position by an arbitrary orthogonal rotation, and then tries by a further series of rotations to get back his original simple structure. The large number of zero loadings in each factor (14 out of 25) means that, after his first choice of reference vectors (provided he is on the right lines), he must always have a large cluster of points round about the origin. This makes it much easier to choose another position for the vectors without losing zero projections or introducing negative projections. So long as the new vector passes through or near to the origin, no

complications arise. When the number of zero projections in each factor is not so large, however, it is not so easy to find suitable further rotations.

It would seem that this will probably be the case in most practical examples. As the labour involved in each rotation is considerable, the position of the axes in the first rotation must be determined with care. In his own illustration Thurstone was concerned only to obtain simple structure. In the present case we have added the further condition that the reference vectors must be capable of psychological interpretation : this often determines their position very effectively, and, in the educational and psychological fields, such a condition would generally apply.

The final step is to bring the vectors back to normal length by dividing each row by the stretching factor which produced E_c from F_0 (Table II). This is most conveniently carried out by multiplying each row of E_1 by the corresponding entry in the first column of F_0.[1]

Table V shows the result of this method of rotation, i.e., it gives the structure on the reference vectors. The method used by Sir Godfrey Thomson (5, p. 257) is slightly different, but yields the same results.

TABLE V. THE STRUCTURE ON THE REFERENCE VECTORS ($F_0\Lambda_{01}$)

	I_1 Number	II_1 Space	III_1 Verbal	IV_1 Schooling	V_1 Reasoning
1	·588	·339	−·013	·446	·001
2	·035	·398	·300	·233	·165
3	·469	·243	−·045	·321	·198
4	·576	·177	·007	·403	·002
5	·174	−·002	·055	·132	·468
6	·467	−·019	·272	·531	−·014
7	·216	·023	·159	·299	·400
8	·635	·105	−·006	·512	·028
9	·241	·025	·106	·357	·452
10	−·001	·057	·246	·276	·571
11	·116	·167	·270	·352	·437
12	·115	·347	·142	·349	·500
13	·203	·520	·082	·443	·483
14	−·001	·395	·185	·109	·231
15	·113	·516	−·004	·257	·549
16	·096	·359	·225	·242	·114
17	−·074	·228	·074	·000	·464
18	·005	·090	·135	·009	·416
19	·045	·215	·068	·000	·386
20	·031	·019	·614	·606	·277
21	·020	·017	·479	·427	·290

Having obtained the equations for the lines (5 in this case) Thomson normalizes them by multiplying the constant term by ·447 ($v + w + x + y + z$) : this is equal to unity for all points on the plane considered. Each of the new reference vectors is thus expressed in terms of the five Landahl axes, which are orthogonal ; and the equations can therefore be normalized directly. In Thurstone's method, on the other hand, the vectors represented by the equations whose coefficients are entered in S_{01} are in each case referred to a frame which is not orthogonal, namely, two Landahl axes and the original centroid axis. In this case, therefore, before we can derive the multiplying factors which will convert the direction numbers into direction cosines, we must refer the new reference vectors to an orthogonal frame. This is done by premultiplying S_{01} by M_0 (the Landahl Matrix) which refers these new reference vectors to the original centroid axes. After normalizing the equations Thomson premultiplies the matrix of coefficients so obtained ($S^1_{01}D^1$) by M_0 which refers the new reference vectors to the original centroid frame, and yields the matrix $\Lambda^1 = M_0S^1_{01}D^1$ which will rotate the original matrix of centroid loadings to their new values.

In Thurstone's method, M_0S_{01} is computed and normalized by post-multiplying it by the diagonal matrix D, giving $\Lambda = M_0S_{01}D$ which is the same as above.

[1] This matrix can also be obtained directly as follows. Form the matrix M_{01} where $M_{01} = L_{01}D_1$. The matrix Λ_{01} which will rotate F_0 into the new reference frame is obtained from M_{01} by leaving out the first column. Then $F_0\Lambda_{01} = E_1$. It is probably best to find E_1 by both methods using one as a check on the other.

The entries in $F_0\Lambda_{01}$ in Table V are the correlations between the tests and the reference vectors, the 5 arbitrary test vectors in terms of which the inter-correlations between the tests can be explained. If we compare Table V with Table III we find a marked similarity, except that there is no general factor in the former. It is possible to identify the primary factors corresponding to vectors I, II, and V as number, space, and reasoning respectively, and those corresponding to vectors III and IV as possibly verbal and schooling factors. If this is correct, then it is evident that the factor called ' verbal ' has a great deal of schooling included in it. While the added freedom in the selection of new positions for the axes is on the whole likely to lead to factors which will be more easily interpreted, it has never been claimed that oblique factors will give a more satisfactory solution in all cases. In this example little seems to have been gained. The factors, though similar, are obviously not the same ; and in our present state of knowledge it is difficult to determine which are nearer to the true factors. If, however, any significance is to be attached to grouping and alignment of test points, then we should have to accept the oblique axes as the better estimate of the true position.

IV. SECOND ORDER FACTORS

When the reference vectors are oblique the primary factors are also oblique ; and the correlations between them can be calculated (Table VI). It will be seen that 4 of the 5 primary factors are positively correlated, i.e., the angles between them are acute, and that the factor is negatively correlated with all the others.

TABLE VI. CORRELATIONS BETWEEN THE PRIMARY FACTORS $(D\Lambda^{-1}\Lambda'^{-1}D_1)$

	I	II	III	IV	V
I	—	·077	·877	—·870	·703
II		—	·311	—·272	·165
III			—	—·898	·609
IV				—	—·600
V					—

An attempt to factorize this matrix by Spearman's method showed that one factor was not sufficient. A centroid analysis was therefore performed. Five factors were obtained with guessed communalities.[1] Two iterations, with improved communalities, furnished four factors. These were then rotated in pairs, until all the loadings (except for factor IV) were positive or zero. These are shown in Table VII.

[1] In a written comment on my paper (for which I am much indebted) Mr. Summerfield enquires why the 5×5 matrix shown in Table VI cannot, by the aid of suitable communalities, be expressed in terms of two factors only. I assumed it would be possible theoretically. But I was unable to reach suitable communalities by iteration. Professor Burt points out that, though 5 variables are needed to determine 2 factors (5, p. 40), at least 3 factors are in general needed to account for 5 variables, and that, on solving the 4×4 minors for the requisite communality, some of the values obtained exceed unity.

TABLE VII. SECOND ORDER FACTORS AFTER ROTATION IN PAIRS (E)

	I′	II′	III′	IV′	S_1	S_2	S_3	S_4	S_5
I	·995	·106	·019	·000	·000				
II	·030	·513	·300	·000		·804			
III	·823	·528	·043	·177			·099		
IV	−·825	−·470	−·019	·177				·258	
V	·713	·000	·470	·000					·520

To obtain the saturations of each of the tests in these 4 second order factors $F_0\Lambda_{01}$ is post-multiplied by $D^{-1}E$. The results of this multiplication [1] are shown in Table VIII.

TABLE VIII. FOUR SECOND ORDER FACTORS ($G = F_0\Lambda_{01}D^{-1}E$)

	I′	II′	III′	IV′	2	3	4	5	h^2	Centroid h^2
1	·736	−·157	·069	·192	·330	−·004	·289	·001	·800	·798
2	·475	·415	·279	·249	·387	·081	·151	·123	·732	·741
3	·783	−·151	·176	·121	·236	−·012	·208	·148	·802	·804
4	·831	−·181	·016	·183	·172	·002	·262	·001	·855	·861
5	·826	−·025	·306	·085	−·002	·015	·086	·350	·914	·923
6	·827	−·105	−·035	·368	−·018	·074	·345	−·010	·957	·970
7	·765	−·044	·271	·210	·022	·043	·194	·299	·834	·847
8	·768	−·356	·066	·225	·102	−·002	·332	·021	·892	·890
9	·650	−·180	·297	·210	·024	·029	·232	·338	·757	·767
10	·566	·064	·423	·242	·055	·067	·179	·427	·784	·788
11	·662	·112	·365	·287	·162	·073	·228	·327	·856	·865
12	·449	·042	·457	·224	·337	·038	·227	·374	·769	·774
13	·359	−·021	·492	·237	·505	·022	·288	·361	·896	·899
14	·438	·383	·317	·138	·384	·050	·071	·173	·643	·646
15	·362	·045	·539	·112	·502	−·001	·167	·411	·885	·886
16	·408	·290	·217	·217	·349	·061	·157	·085	·503	·507
17	·440	·226	·409	·036	·222	·020	·000	·347	·583	·592
18	·729	·241	·329	·069	·087	·037	·006	·311	·809	·826
19	·685	·245	·345	·033	·209	·018	·000	·289	·777	·794
20	·497	·190	·236	·567	·018	·166	·393	·207	·885	·887
21	·546	·203	·258	·422	·017	·130	·277	·217	·725	·722

This shows loadings on four fairly general second order factors together with four of the five factors from Table V with loadings proportional to the values in the table. One primary factor has disappeared : the whole of its variance is accounted for by the four second order factors. Since the factors are now orthogonal, they should, and do (as closely as can reasonably be expected), yield the original communalities and correlations.[2] Of the second order factors Factor I may be interpreted as a general factor, weighted on the arithmetical side, Factor III as a reasoning factor, and Factor IV, which has more or less absorbed primary factor 3, as a verbal factor. Factor II has negative loadings in all of the arithmetic tests except No. 10, and positive loadings on the others (except No. 13) : it may be interpreted as an interest or novelty factor.

[1] The case in which (i) complete simple structure is revealed in the test configuration and (ii) the second order correlation matrix is of unit rank is discussed by Thurstone (4, p. 422) and Thomson (5, p. 297).

[2] Agreement with the original communalities and correlations is obtained only if the correlations between the primary factors are analysed completely. In one case a significance test indicated that the residuals after the extraction of the second factor were not significant. Using this analysis, and obtaining 2 second order factors and 5 primaries, we sometimes get communalities greater than unity.

The analysis into oblique factors followed by the extraction of second order orthogonal factors has therefore yielded 7 distinct orthogonal factors, general intelligence, interest, number, space, verbal, reasoning, and schooling. (There were 8 factors but 2 were interpreted as reasoning.) The orthogona rotation in pairs of the same centroid loadings produced 5 factors, general intelligence, number, space, verbal, and reasoning, which (though not quite the same as the 5 of that name in the 7 above) were very similar to them.

V. CONCLUSIONS

What has been gained by this highly detailed analysis ? Five centroid factors were rotated in pairs, and yielded 5 fairly clearly defined orthogonal factors. Since most human attributes and traits, especially the mental qualities named and defined, are in fact correlated, a better explanation of the correlations was sought in terms of correlated factors. These, owing to the conditions of the rotation, could not include a general factor ; but, since they were correlated, could presumably be explained in terms of a common factor. When the correlations between them were analysed, there was not one common factor but 4. The number is thus increased to 8 in all, one of the original 5 being completely absorbed in the 4 secondary factors. Evidently then each of the 5 oblique factors is highly complex, being composed of 4 second order factors and a part specific to itself. During the investigation of which the analysis is part, four complete analyses were carried out ; in each case at least 3 second order orthogonal factors were needed to account for the correlations between the primary oblique factors. It therefore appears certain that the original orthogonal factors, though independent, are by no means simple unitary factors ; and one is driven to conclude that the second order factors are probably complex also. This is in agreement with the views expressed by Thomson in his presidential address to the Psychology Section of the British Association. Factors, he said, were in his opinion " not at all unitary, and only partially innate, being to a very considerable extent built up in the individual mind by many external influences of neighbourhood, school, various experiences and opportunities, availability of books and tools, and such like." If so, then more detailed analyses ought to yield more factors, each factor becoming increasingly unitary ; and theoretically, could a suitable battery of tests be devised, one might finally arrive at an analysis into the simple unitary factors.

With most empirical data, however, this is impossible owing to the effect of error or other uncontrollable influences ; and it seems to the present writer that considerable caution must be exercised in analysing data by the above method. We start with fallible data subject to sampling and other errors ; and with every further subdivision of the communalities we get factors with loadings which approach closer and closer to the borderline of non-significance. A time comes when a halt must be called. In the analysis here described it is questionable whether one is justified in going beyond the original 5 orthogonal or oblique factors, although the analysis into second order factors yielded at least one interesting bipolar factor.

VI. SUMMARY

1. Inter-correlations, obtained from 21 tests applied to 136 boys, were factorized by Thurstone's centroid method.

2. The 5 factors so extracted were rotated (*a*) in pairs and (*b*) by the method of extended vectors. In both cases, number, space, reasoning, and verbal factors were identified. The orthogonal rotations produced a general factor, and the oblique rotations produced a schooling factor.

3. The correlations between the 5 oblique primary factors were analysed, and furnished 4 second order orthogonal factors, identified as general intelligence, reasoning, verbal, and interest factors.

4. However, it is concluded that, in most practical experiments, it is hardly justifiable to go beyond the original orthogonal or oblique factors.

REFERENCES

1. Coombs, C. H. (1941). ' Criterion for significant common factor variance.' *Psychometrika*, VI, 267–72.
2. McNemar, Q. (1942). ' On the number of factors.' *Psychometrika*, VII, 9–18.
3. Lawley, D. N. (1943). 'The application of the maximum likelihood method to factor analysis.' *Brit. J. Psych.* XXXIII, p. 172.
4. Thurstone, L. L. (1947). *Multiple Factor Analysis*. Chicago : The University of Chicago Press.
5. Thomson, G. H. (1946). *The Factorial Analysis of Human Ability*. (2nd ed.) London : University of London Press.

Vol. IV
Part I

The British Journal of Psychology,
Statistical Section

March,
1951

THE NUMERICAL SOLUTION OF LINEAR EQUATIONS[1]

By CYRIL BURT

University College, London

I. *Chief Types of Problem.* II. *General Method of Solution.* III. *Routine Methods.* IV. *The Doolittle Method.* V. *Inverses.* VI. *Triangular Group Factor Methods.* VII. *Calculation of Determinants.* VIII. *Partial Correlations.* IX. *Summary.*

I. CHIEF TYPES OF PROBLEM

General and Special Cases. In the work of the statistical psychologist one of the most frequent computational problems is the solution of a system of simultaneous linear equations. The general type of problem may be most conveniently expressed in matrix notation as follows :

" Given $AX = C$: required to find X." (i)

The following are the commoner cases in which such an equation occurs :

(i) *Multiple Correlation :* in this case A will denote the $n \times n$ symmetric matrix formed by the inter-correlations between the tests, C will be a column-vector containing the correlations of the tests with the criterion, and X the column-vector of partial regressions we are required to find.

(ii) *Factor Measurements :* here C will represent the $n \times r$ matrix of factor saturations, and X the matrix of weights for computing the factor measurements.

(iii) *Least Squares.* Both the foregoing cases are really special applications of a more general procedure. Given a number of fallible and discrepant observations for a single set of values (say, the results of n intelligence tests applied to a sample of N persons), it is required to deduce a method for combining these observations so as to obtain the best possible estimates for the quality they are intended to measure. The classical method of treating such a problem is to apply the principle of least squares. With that procedure equation (i) will represent the so-called ' normal equations ' ; A will be a symmetric matrix of product sums ; and X the column-vector of weights, which are to be so determined as to yield the closest fit.

(iv) *Inverses.* In calculating an inverse we must take C to be the identity matrix. Thus, when found, $X = A^{-1}$. For the rest the working procedure will be much the same as before.

(v) *Determinants.* Many problems in statistical psychology (e.g., partial correlation) call for the calculation of determinants ; and the procedure devised for solving equation (i) can be adapted for this purpose.

(vi) *Factor Saturations.* To compute the factor saturations for a given table of correlations, covariances, product sums, or the like, we have to solve, not linear equations, but a matrix equation of the second degree. In this case A will stand for the transpose of X ; and we shall in effect be required to solve $X'X = C$, where C is symmetric and the columns of X' may be called its components or ' factors.' As has long been known in matrix algebra, a preliminary solution of what I will call the ' factorial problem ' often greatly facilitates the final or approximate solution of equations such as those described above.

Working Methods. The methods of solving systems of linear equations are of three main types : (i) algebraic solutions given in books on theory ; (ii) practical

[1] At a recent meeting of the Council of the British Psychological Society it was suggested that, in future numbers of the *Journal*, the editors should publish from time to time comparatively elementary articles of an expository type on the statistical problems more commonly encountered by the research student in psychology. In reply to requests addressed to a large group of readers, the questions discussed in the following pages were among those most frequently mentioned.

solutions for computers, nearly all of them deriving ultimately from the procedure suggested by Gauss ; (iii) new methods specially developed for computers using a modern calculating machine. Most manuals of statistical psychology describe a working procedure, usually of the second type. Those generally recommended (e.g., Doolittle's or Aitken's) are somewhat complicated for the beginner to follow ; and, as a rule, no attempt is made to explain the underlying algebraic principles. My own experience suggests that such an explanation makes the arithmetical computations simpler to understand, easier to remember, and more interesting to apply. In this paper part of my aim will be to show how formulæ and procedures, which look somewhat forbidding in matrix notation, are merely generalizations of elementary algebraic principles with which nearly every student will be familiar.

To understand the essential portions of the exposition outlined here, the junior student will need little more than an acquaintance with the principles of substitution and elimination, as described in any elementary text-book of algebra [1] ; the more advanced will find it helpful to have a rudimentary knowledge of determinants and matrix notation.

Historical Note. The general rule for solving linear equations, which goes by his name, was given by G. Cramer in 1750. However, in statistical work the numerical coefficients attached to the unknowns generally form a symmetric matrix ; and this enables somewhat simpler procedures to be employed.

The practical treatment of the question virtually begins with Lagrange's [2] method for reducing a quadratic form to a sum of squares by means of a ' rational congruent ' transformation. If $x'Ax$ is the quadratic form, then we find, by progressive steps, a transformation $y = Tx$, such that $x'Ax = y'Vy$, where $V = T'AT$ is a diagonal matrix and T is a *triangular* matrix.[3] The transformation can be carried out by a terminating sequence of ' elementary ' operations, each of which involves purely *rational* processes only. However, the congruent reduction to a ' rational canonical form ' is not unique. The older [4] orthogonal reduction to the ' classical canonical form ' (the transformation to a diagonal matrix of latent roots adopted in Pearson's analysis into ' principal components ') *is* unique ; but in general it entails irrational operations, since it implies solving the ' characteristic equation ' of A, viz., $|A - \lambda I| = 0$.

Gauss's method for solving the ' normal equations ' [5] is a development of Lagrange's procedure. In English the clearest explanation of the actual working is that of W. W. Johnson.[6] Johnson's mode of tabulating the figures shows very plainly how Gauss's method really involves a transformation by a triangular matrix ; i.e., as he puts it, " in the determinant all the elements below the principal diagonal vanish " (p. 135 ; cf. eq. 8, p. 123, and tables on pp. 140–2) ; and, partly for that reason, I have adopted it here.

The relation between the theory of the product-moment correlation and the method of least squares is brought out in early papers on the subject by Yule (at that time assistant in Karl Pearson's department).[7] He shows how a set of n correlated deviations, $x_1, x_2, \ldots, x_n$, may be reduced to terms of n ' uncorrelated deviations,' $x_1, x_{2\cdot1}, x_{n\cdot1} \ldots, x_{(n-1)}$ by what is essentially Lagrange's procedure. This naturally suggested the use of these ' uncorrelated deviations ' as ' factors ' in psychological work (see p. 42).

The earliest and most detailed treatment of such problems by the aid of matrix notation is

[1] E.g., Fine, H. B., *College Algebra*, ch. IV, p. 131 ; for determinants, see *ibid.*, ch. XXXI, pp. 492f.

[2] Lagrange, J. L. (1759), *Œuvres* (1867 ed.), I, pp. 3–20. Lagrange's reduction may be regarded as a generalization of the familiar process of factoring a quadratic expression by ' completing the square ' (cf. *College Algebra*, p. 187) : it is this, as we shall see, that is responsible both for the ' hierarchies ' of coefficients that have to be subtracted and for the square roots that appear in the factors.

[3] Bôcher, M., *Higher Algebra*, 1907, p. 131. Aitken, A. C., *Theory of Canonical Matrices*, 1932, p. 85, theorem I.

[4] Aitken, A. C., *loc. cit.*, p. 58 : (see also pp. 10f. and 33f. for definitions of the technical expressions used in this paragraph).

[5] Gauss, C. F. (1823), *Theoria combinationis observationum erroribus minimus obnoxiæ* : *Supplementum* (*Werke*, Bd. IV, pp. 27f.). The commoner accounts seem to be derived from the somewhat fuller exposition of Gauss's working method given by Encke, *Berlin, Astron. Jahrb.* (1835), pp. 267f.

[6] *The Theory of Errors and Method of Least Squares*, 1892, ch. IX, ' Gauss's Method of Substitution.' (This little volume was a favourite manual with Galton and K. Pearson ; and the procedure there described was commonly adopted by their early students in the Biometric Laboratory at University College : my own mode of tabulation is taken from Johnson.)

[7] Yule, G. U., ' On the Theory of Correlation,' *J. Roy. Stat. Soc.*, LX, 1897, p. 849 ; ' On the Theory of Correlation for any Number of Variables Treated by a New System of Notation,' *Proc. Roy. Soc.*, A, LXXIX, 1907, pp. 182f. *Id.*, *Introduction to the Theory of Statistics*, 1912, p. 331.

contained in Cullis's three-volume work on *Matrices and Determinants*.[1] He distinguishes clearly between 'rational' methods of reduction (p. 280), such as that due to Lagrange (p. 290), and 'unrestricted' modes of reduction involving irrational operations such as taking square roots (p. 293). More especially he points out that, if $A = XX'$ and is thus a symmetric matrix, "a particular solution for X can *best* be obtained" by a procedure of the latter type: we first form the $1 \times n$ vector

$$h' = \frac{1}{\sqrt{a_{11}}} \left[a_{11},\ a_{12},\ \ldots\ a_{1n} \right]; \qquad \text{(ii)}$$

we then have

$$A = hh' + \frac{1}{a_{11}} \begin{bmatrix} 0, & 0 \\ 0, & B \end{bmatrix}. \qquad \text{(iii)}$$

B in turn can be treated in the same way, and so we continue; we thus finally obtain X in triangular form (p. 357; cf. also p. 358, case I).

For the part played by such reductions in simplifying and evaluating the expression for the normal multivariate distribution, see Wilks, S. S., *Mathematical Statistics*, 1947, pp. 63–8.[2]

II. GENERAL METHOD OF SOLUTION

The Chief Methods Available. Let the given equation be $Ax = c$, that is,

$$\begin{aligned} a_{11}x_1 + \ldots + a_{1n}x_n &= c_1 \\ \ldots \quad \ldots \quad \ldots \quad \ldots & \\ a_{n1}x_1 + \ldots + a_{nn}x_n &= c_n. \end{aligned} \qquad \text{(i}b\text{)}$$

Then, for purposes of practical computation, there are three main ways of carrying out the requisite calculations:

(i) *Multiplication by Inverse.* Assuming $|A| \neq 0$, we can pre-multiply both sides of the matrix equation by A^{-1}; we then obtain

$$x = A^{-1}c, \qquad \text{(iv)}$$

and, since A^{-1} is unique, this is the only solution. In practice, therefore, we may compute the inverse, A^{-1}, and then post-multiply the result by the vector c. Not infrequently A is singular; but in such cases it may be possible to subject it to a preliminary transformation which will convert it into a non-singular matrix of lower order.[3]

(ii) *Cramer's Rule.* If we express the above solution in determinantal form, we have

$$x_1 = \frac{A_1}{|A|}, \qquad \ldots, \qquad x_n = \frac{A_n}{|A|}, \qquad \text{(v)}$$

[1] Cullis, C. E., *Matrices and Determinants*, II, 1918, pp. 280–93, 357–60.

[2] It will be noted that Wilks uses a square root transformation to reduce the quadratic form to a sum of squares (eq. *c*, p. 66); he also notes that the evaluation of determinants by 'pivotal condensation' is essentially Lagrange's method (p. 65). Kendall adopts Yule's notation and reduction: *Advanced Theory of Statistics*, I, pp. 376–8.

[3] In psychological work two main cases arise. (i) The matrix A may be, not square, but rectangular: e.g., it may be a measurement matrix with more columns than rows. It can then generally be reduced to a non-singular square matrix either (*a*) by self-multiplication (e.g., taking $AA' = R$), or (*b*) by selection of linearly independent columns (cf. Burt, C., 'Correlations between Persons,' *Brit. J. Psych.*, XVIII, 1937, pp. 85f., eq. (36)). (ii) Though square, the matrix A may be 'doubly-centred': e.g., it may be a correlation matrix with both rows and columns adding up to zero. This is the case with residual matrices on which bipolar factors are based; but it may also result from the conditions of the experiment or the marking: (cf. *id.*, *Brit. J. Med. Psychol.*, XVII, pp. 158–88). In such a case the singular matrix can be reduced to one of lower order by using what I have called a 'trans-classificatory matrix.' When complete (i.e., square) such a matrix is not only orthogonal but also (as I have called it) 'equilibrated' (i.e., all its rows and/or columns sum to zero, except the first). Two forms are especially useful: (i) the 'solid' form, consisting of equal elements and a balanced sign pattern (both rows and columns summing to zero); (ii) the 'triangular' form, consisting of unequal elements and unbalanced sign pattern (only columns summing to zero): for examples, see this *Journal*, II, p. 112, Table VC and VB. To obtain 'reduction matrices' the first column must be omitted (or treated as inoperative). To reduce a non-singular group factor matrix to bipolar form, we must post-multiply it by the reduction matrix; and to reduce a doubly-centred correlation matrix to singular form, we must pre- and post-multiply by the reduction matrix.

where A_i denotes the n-rowed determinant obtained by substituting the column of elements $c_1, \ldots c_n$ for the elements of the ith column of $|A|$, e.g., if $n = 3$,

$$x_3 = \frac{\begin{vmatrix} a_{11} & a_{12} & c_1 \\ a_{21} & a_{22} & c_2 \\ a_{31} & a_{32} & c_3 \end{vmatrix}}{\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}} \qquad \text{(vi)}$$

We may therefore proceed by evaluating the $(n + 1)$ determinants in equation (v), and then calculating such quotients as that shown in equation (vi). In this country the best known method of evaluating determinants is that called ' pivotal condensation,' generally used in the form described by Aitken (cf. also *College Algebra*, 1904, p. 917). Aitken himself points out that practically all the condensation methods in actual use " are variants of a method commonly ascribed to Chiò (1853), but virtually used by Gauss more than forty years earlier." [1]

(iii) *Gauss's Method.* Gauss's own procedure, as applied to problems of least squares, includes two main stages. (*a*) It begins with a ' forward solution,' consisting of a sequence of ' elimination equations.' By taking (say) the first of our given equations we can express x_1 in terms of the $(n - 1)$ other unknown quantities ; this expression is then substituted in the remaining equations, with the result that x_1 is completely eliminated. Next, taking the second equation (as thus modified), x_2 is eliminated in precisely the same way ; and the process is repeated, until we are left with a single equation which yields a numerical value for x_n. (*b*) We then turn to the ' back solution,' which consists of a sequence of ' substitution equations.' The value found for x_n is substituted in the last equation but one, and we thus find x_{n-1}. Both are then substituted in the equation for x_{n-2}. And so, working backward step by step, we eventually obtain values for all the n unknowns. With slight modifications this is the basis of the procedures suggested here.

Regression Equations. In statistical psychology the most frequent, and in many ways the most typical, form of the general problem consists in the calculation of partial regressions, i.e., of ' weights,' for estimating an external or internal ' criterion ' by combining a number of observed measurements. This requirement occurs both in multiple correlation and in factor analysis. We may therefore choose it for our first illustration.

Example.[2] The procedure proposed can best be grasped by considering a concrete case. Let it be required to solve the following equations :

$$\begin{aligned} 1{\cdot}0x_1 + {\cdot}6x_2 + {\cdot}5x_3 + {\cdot}3x_4 &= {\cdot}2 && (a.\text{i}) \\ {\cdot}6x_1 + 1{\cdot}0x_2 + {\cdot}7x_3 + {\cdot}5x_4 &= {\cdot}4 && (a.\text{ii}) \\ {\cdot}5x_1 + {\cdot}7x_2 + 1{\cdot}0x_3 + {\cdot}6x_4 &= {\cdot}6 && (a.\text{iii}) \\ {\cdot}3x_1 + {\cdot}5x_2 + {\cdot}6x_3 + 1{\cdot}0x_4 &= {\cdot}8 && (a.\text{iv}) \end{aligned}$$

The figures on the left, we may suppose, come from a table of inter-correlations between four intelligence tests, while the figures on the right represent the correlations of the tests with an external criterion. The unknown ' weights,' x_1, x_2, x_3, and x_4, will be the partial regressions.

(a) *Forward Solution (by Elimination).* Using the first equation (*a*.i) we have at once

$$x_1 = -{\cdot}6x_2 - {\cdot}5x_3 - {\cdot}3x_4 + {\cdot}2 \qquad (a.\text{i bis})$$

This can now be used to eliminate x_1 from all the other equations. To eliminate it from the second equation (*a*.ii), we must first multiply x_1 by its coefficient in that equation, and so obtain from (*a*.i bis)

$${\cdot}6x_1 = -{\cdot}36x_2 - {\cdot}30x_3 - {\cdot}18x_4 + {\cdot}12 .$$

[1] *Determinants and Matrices*, 1939, p. 46. The method described in most English school books appears to be based on the form developed by C. L. Dodgson (Lewis Carroll) in his *Elementary Treatise on Determinants* (1876).

[2] The example, and the discussion which follows, is taken from a War Office memorandum on ' The Solution of Simultaneous Linear Equations,' which I prepared for the Directorate of Selection of Personnel in 1942. I am grateful for permission to publish it. The ' correlations ' are, of course, quite fictitious and are chosen so that the reader can do most of the illustrative calculations in his head.

On substituting the expression thus obtained, equation (*a*.ii) becomes

$$(1{\cdot}00 - {\cdot}36)x_2 + ({\cdot}70 - {\cdot}30)x_3 + ({\cdot}50 - {\cdot}18)x_4 = {\cdot}40 - {\cdot}12.$$

We can make similar substitutions in the two remaining equations. Evidently the whole process is equivalent to subtracting the four following equations from the initial four.

(1·00)	$1{\cdot}00x_1 + {\cdot}60x_2 + {\cdot}50x_3 + {\cdot}30x_4 = {\cdot}20$	(*b*.i)
(·60)	$\cdot 60x_1 + {\cdot}36x_2 + {\cdot}30x_3 + {\cdot}18x_4 = {\cdot}12$	(*b*.ii)
(·50)	$\cdot 50x_1 + {\cdot}30x_2 + {\cdot}25x_3 + {\cdot}15x_4 = {\cdot}10$	(*b*.iii)
(·30)	$\cdot 30x_1 + {\cdot}18x_2 + {\cdot}15x_3 + {\cdot}09x_4 = {\cdot}06$	(*b*.iv)

These equations are derived from equation (*a*.i) by multiplying it throughout by the coefficients of x_1 in (*a*.i–iv), here shown in brackets on the left. They in turn correspond to the column-vector, h, in Cullis's equation (eq. (ii), p. 33, above) and may be regarded as a column of saturations for the first factor : the hierarchy of figures so obtained corresponds to the expression that he writes as hh' (eq. (iii), p. 33). On subtracting this ' hierarchy,' we are left with the residuals $A - hh' = B_0$ (where B_0 consists of an $(n-1) \times (n-1)$ matrix bordered by a row and column of zeros), or, in explicit form (omitting the zeros) :

$\cdot 64x_2 + {\cdot}40x_3 + {\cdot}32x_4 = {\cdot}28$	(*c*.ii)
$\cdot 40x_2 + {\cdot}75x_3 + {\cdot}45x_4 = {\cdot}50$	(*c*.iii)
$\cdot 32x_2 + {\cdot}45x_3 + {\cdot}91x_4 = {\cdot}74$	(*c*.iv)

We note that, as a result of the subtraction, x_1 is now entirely eliminated.

We next eliminate x_2 in precisely the same way. The equations to be subtracted are

·80 (1·000)	$\cdot 64x_2 + {\cdot}40x_3 + {\cdot}32x_4 = {\cdot}280$	(*d*.ii)
·50 (·625)	$\cdot 40x_2 + {\cdot}25x_3 + {\cdot}20x_4 = {\cdot}175$	(*d*.iii)
·40 (·500)	$\cdot 32x_2 + {\cdot}20x_3 + {\cdot}16x_4 = {\cdot}140$	(*d*.iv)

Once again the figures can be directly calculated by the formula quoted from Cullis. But now the leading term in (*c*.ii) is no longer unity, but ·64. In accordance with his equation therefore we form a second vector of factor saturations, h_2', by dividing the top row of residuals (*c*.ii) by the square root of the leading term ($\sqrt{\cdot 64} = {\cdot}80$). This is equivalent to reducing all the quantities in that row in the same proportion. We thus obtain

$$\frac{\cdot 64}{\sqrt{\cdot 64}} = {\cdot}80\,; \quad \frac{\cdot 40}{\sqrt{\cdot 64}} = {\cdot}50\,; \quad \frac{\cdot 32}{\sqrt{\cdot 64}} = {\cdot}40.$$

(These saturations are entered on the extreme left of equations (*d*).) A fresh hierarchy, $h_2'h_2$, is then constructed from these figures by multiplying each by each in the usual way.

Now, with an empirical table, the explicit calculation of the square root may be a little troublesome ; and, for the purposes of eliminating x_2, all that is necessary is to reduce the figures in the several rows in the same proportion, viz., ·64 : ·40 : ·32 = 1·000 : ·625 : ·500. These proportional values form an alternative set of multipliers, and are given in brackets at the left. For reasons that will be clear in a moment I shall call them ' unstandardized ' factor saturations. Each is now used to multiply the original coefficients *as they stand*. We thus avoid the irrational operation of finding a square root ; and secure precisely the same hierarchy of figures as before.

On subtracting this second hierarchy from the residuals (c.ii, iii, and iv) we obtain

$\cdot 50x_3 + {\cdot}25x_4 = {\cdot}325$	(*e*.iii)
$\cdot 25x_3 + {\cdot}75x_4 = {\cdot}600$	(*e*.iv)

x_2 has now been eliminated. The final hierarchy can be constructed in exactly the same way as the previous hierarchies. We obtain

·707 (1·00)	$\cdot 500x_3 + {\cdot}250x_4 = {\cdot}3250$	(*f*.iii)
·354 (·50)	$\cdot 250x_3 + {\cdot}125x_4 = {\cdot}1625$	(*f*.iv)

The last residual is therefore

$$\cdot 625x_4 = {\cdot}4375. \qquad (g.\text{iv})$$

All the unknowns except the last have now been eliminated ; and we have at once

$$x_4 = \frac{\cdot 4375}{\cdot 625} = {\cdot}70. \qquad (h.\text{iv})$$

(b) *Backward Solution (by Substitution).* The values for the three other weights can be calculated by successive substitution. Inserting the value just obtained for x_4 in equation (*f*.iv), we have

$$\cdot 250x_3 + {\cdot}125 \times {\cdot}70 = {\cdot}1625.$$

Hence $\cdot 250x_3 = {\cdot}0750$, and $x_3 = {\cdot}30.$ (*h*.iii)

Substituting the values for x_4 and x_3 in equation (*d*.iv) we have

$$\cdot 32x_2 + {\cdot}20 \times {\cdot}30 + {\cdot}16 \times {\cdot}70 = {\cdot}140\,.$$

Hence $\cdot 32x_2 = -{\cdot}032$, and $x_2 = -{\cdot}10.$ (*h*.ii)

Finally, substituting for x_4, x_3, and x_2 in (*a*.iv) we have

$$\cdot 3x_1 + \cdot 5 \times (-\cdot 10) + \cdot 6 \times \cdot 30 + 1\cdot 0 \times \cdot 70 = \cdot 8.$$

Hence $$\cdot 3x_1 = -\cdot 03, \text{ and } x_1 = -\cdot 10 . \qquad (h.\text{i})$$

This completes the solution for the regression coefficients.

General Principles. The method I have described may be called ' the method of hierarchical subtraction.' It consists of a repeated cycle of processes : (i) the calculation of the factor saturations or their equivalent by *division ;* (ii) the calculation of a hierarchy by *multiplication ;* (iii) the calculation of the residuals by *subtraction.* Since the matrices to be subtracted, like all correlational hierarchies, are symmetric, it has the great advantage of preserving the symmetry in the successive sets of coefficients. It thus diminishes the number of incidental figures which have to be computed, or alternatively provides a helpful check.

Rational and Non-rational Methods. It will be noted that with the method as I have so far described it, each hierarchy can be calculated in one of two ways. In either case it is always derived by pre-multiplying a *row* of figures by a *column.* But with the first method *both* row and column are obtained by dividing a vector of residuals. This may be called a ' method of double division ' (though, with a symmetric matrix, the division is only performed once, since, not only the vectors, but also the divisors, are the same, viz., the square root of the first figure in the residual row). With the second method the *row* of residuals may be left as it is without division, and the *column* obtained by dividing that row by its own leading term just as it stands. This may be called a ' method of single division.' [1] The first method of reduction, it will be observed, involves extracting a square root at each successive stage ; it is therefore ' irrational.' The second can be carried out by purely ' rational ' operations only.

Now in factor analysis we require to calculate *standardized* factor saturations ; and this necessarily involves division by the square root of the factor variance.[2] Hence the former or ' irrational ' method is unavoidable in factorial work ; moreover, as we shall find later on, it is generally the most convenient for calculating inverses and determinants. But for the calculation of regressions and multiple correlations the second or ' rational ' method is, in my view, by far the simplest and most expeditious.[3]

The Reduction of a Quadratic Form. The underlying reasons for these various computational steps will become clearer if we express the processes in symbolic form, and relate them to analogous processes arising in more familiar types of problem. As we have seen, the method is, in effect, a special application of the traditional mode of reducing a ' quadratic form ' to its normal or ' canonical ' shape, i.e., to a sum of squares ; and this in turn is merely a generalization of simple algebraic processes with which every schoolboy is familiar.

A polynomial, all of whose terms are of the same degree, is called a ' form.' If the degree is 2 it is called a ' quadratic form.' Such a form can always be reduced to a sum of squares by a non-singular linear transformation which will abolish the product term (xy). Thus, to take the simplest possible case, $4x^2 + 12xy + 25y^2 = (2x + 3y)^2 + (4y)^2 = X^2 + Y^2$, where $X = 2x + 3y$ and $Y = 4y$, so that here the factor saturations would be [2, 3] for factor I, and [0, 4] for factor II. And generally the binary ' form ' $ax^2 + 2hxy + by^2$ can always be reduced to canonical shape by first ' completing the square ' for the first two terms, and then expressing the remainder as a second square, thus : $\left\{\frac{1}{\sqrt{a}}(ax + hy)\right\}^2 + \left\{\frac{1}{\sqrt{a}}\sqrt{ab - h^2}.y\right\}^2$. Note that it is the process of ' completing the square ' that gives rise to the hierarchy which has to be subtracted, viz., $\begin{bmatrix} a^2x^2 + ahxy \\ + ahxy + h^2y^2 \end{bmatrix}$, or, to take the

[1] There is also a ' method of multiple or complete division ' in which the numerical coefficient of x_1 is converted to unity in *every* equation. To do this *each* row must be divided by *its own leading term ;* one row, usually the first, is then subtracted from the others : (cf. *College Algebra*, p. 132, and Huntington, E. V., *Handbook of Mathematical Statistics*, ch. iv). For a beginner this is perhaps the simplest method of all. But it no longer preserves the symmetry of the successive sets of coefficients.

[2] See ' Influence of Differential Weighting,' this *Journal*, II, p. 115, equations (xiii) and (xiv).

[3] However, the former has certain advantages for the solution of many least squares problems. As so applied, I have commonly referred to it as the ' factorial method.' It has been widely used on the Continent : it forms, for example, the basis of the so-called ' Cracowian algorithm ' for solving least squares equations (cf. *Astron. Journ.*, L, 1942, pp. 38–41). Dwyer has recently proposed to reintroduce it for determining partial regressions and multiple correlations : his basic equation is apparently the same as that given by me in *Factors of the Mind*, p. 307 (see Dwyer, P. S., ' The Square Root Method and its Use in Correlation and Regression,' *J. Amer. Stat. Assoc.*, XL, 1945, pp. 493–503).

numerical example, $\begin{bmatrix} 4x^2 + 6xy \\ + 6xy + 9y^2 \end{bmatrix}$. The standardized factor saturations are thus $\left[\frac{1}{\sqrt{a}}a, \frac{1}{\sqrt{a}}h\right]$ and $\left[0, \frac{\sqrt{ab-h^2}}{\sqrt{a}}\right]$. But, if we place the expression $\frac{1}{\sqrt{a}}$ outside the curled brackets when completing the square, we can avoid the square root. We also observe that, unless the remainder-term is non-negative, i.e., unless the 'discriminant' $\begin{vmatrix} a & h \\ h & b \end{vmatrix} = (ab - h^2)$ is positive or zero, the saturation for the second factor will not be real.[1]

Let us now generalize this procedure to the case of the n-ary quadratic form $\sum_i\sum_j a_{ij}x_ix_j = x'Ax$, where A will be symmetric. All that is necessary is to repeat, as often as is necessary, the double process of completing the square and subtracting the hierarchy of coefficients so formed. The operation itself can readily be expressed by matrix multiplication. Taking the example given above, the first step in the reduction is equivalent to pre- and post-multiplying the given matrix A by a unimodular transformation matrix, T_1, and its transpose, T_1', where T_1 will consist of the 'hollow triangle' formed by a unit matrix with a negative 'leg,' viz.:

$$\begin{bmatrix} 1{\cdot}0 & 0 & 0 & 0 \\ -{\cdot}6 & 1{\cdot}0 & 0 & 0 \\ -{\cdot}5 & 0 & 1{\cdot}0 & 0 \\ -{\cdot}3 & 0 & 0 & 1{\cdot}0 \end{bmatrix}$$

For the second step a negative leg will appear in the second column; and so on. Continuing in this way, we finally obtain $T_3T_2T_1AT_1'T_2'T_3' = TAT' = V$, *a diagonal matrix*,[2] where (as the reader can easily verify)

$$T = \begin{bmatrix} 1{\cdot}0000 & {\cdot}0000 & {\cdot}0000 & {\cdot}0000 \\ -{\cdot}6000 & 1{\cdot}0000 & {\cdot}0000 & {\cdot}0000 \\ -{\cdot}1250 & -{\cdot}6250 & 1{\cdot}0000 & {\cdot}0000 \\ +{\cdot}0625 & -{\cdot}1875 & -{\cdot}5000 & 1{\cdot}0000 \end{bmatrix} \text{ and } V = \begin{bmatrix} 1{\cdot}000 & {\cdot}000 & {\cdot}000 & {\cdot}000 \\ {\cdot}000 & {\cdot}640 & {\cdot}000 & {\cdot}000 \\ {\cdot}000 & {\cdot}000 & {\cdot}500 & {\cdot}000 \\ {\cdot}000 & {\cdot}000 & {\cdot}000 & {\cdot}625 \end{bmatrix}$$

Since T is a triangular matrix with units in the leading diagonal, $|T| = |T'| = 1$; and hence $|A| = |V| = v_1v_2 \ldots v_r$, where r is the rank of A (see p. 32).

It is easy to relate this process to the factorization of a table of correlations. Since A here represents the correlation matrix, let us adopt the usual notation, and write $A = R = FF'$, where F denotes the *standardized* factor saturations (with square-sums of rows all = 1). Then, as the reader may verify for himself, the matrix of *unstandardized* factor saturations, F_u say (the bracketed 'multipliers' in the preceding example), will be the inverse of T; i.e., $T^{-1} = F_u = FV^{-\frac{1}{2}}$, and therefore $T = V^{\frac{1}{2}}F^{-1}$. We thus obtain $TRT' = V^{\frac{1}{2}}F^{-1} \cdot FF' \cdot F'^{-1}V^{\frac{1}{2}} = V$, where $v_1, v_2, \ldots v_n$ are the squares of the diagonal elements of F. (For F and F^{-1}, see Table III.)

III. ROUTINE METHODS

Method of Tabulation. In undertaking practical computations it will obviously be unnecessary to write out all the equations explicitly at each successive stage; only the actual figures need be entered in the tables. And, since each square matrix of coefficients and residuals is 'symmetric,' only the upper triangle need be calculated and written down. The subtractions will be easier to carry out if (following Johnson) the initial matrix of correlations (and each subsequent matrix of residuals) is written with a blank line beneath each row, and the hierarchical figures to be subtracted are then entered (as each is calculated) in the blank space so left.

If the coefficient of the multiple correlation is required as well as the partial regressions, the simplest plan will be to treat the column of criterion correlations as making (with the inter-correlations) a single symmetric matrix, thus including an $(n + 1)$th row as well as an $(n + 1)$th column. With these modifications the work sheet will appear as in Table I.[3]

[1] Cf. *College Algebra*, pp. 187–9 and 308. The student may recall the condition that the equation $ax^2 + 2hxy + by^2 = 1$ shall be the equation of an ellipse, viz., that the discriminant $ab - h^2$ must be positive. He will also find it instructive to compare with the foregoing *congruent* transformation the *orthogonal* transformation (or 'rotation') whereby such an ellipse is referred to its principal axes (cf. the procedure used in discussing the normal correlation surface, Yule and Kendall, pp. 230f.).

[2] The whole procedure expressed in symbolic form will be found in any book on matrix algebra; e.g., Aitken, *loc. cit.*, p. 85, Frazer *et al.*, *loc. cit.*, p. 30.

[3] The table is taken from the earlier memorandum, already cited. An abridged account will be found in Vernon, P. E., *Notes on Statistical Methods*, 1945, pp. 6–11.

TABLE I. MULTIPLE CORRELATION AND PARTIAL REGRESSIONS

A. *Forward Solution*

Multipliers					
	1·00	·60	·50	·30	·20
1·00	1·00	·60	·50	·30	·20
		1·00	·70	·50	·40
·60	.....	·36	·30	·18	·12
			1·00	·60	·60
·50	.....	.....	·25	·15	·10
				1·00	·80
·30	.....	.....	.....	·09	·06
					1·00
·20	.....	.....	.....	.....	·04
		·64	·40	·32	·28
1·0000	.....	·640	·400	·320	·280
			·75	·45	·50
·6250	.....	.....	·250	·200	·175
				·91	·74
·5000	.....	.....	.....	·160	·140
					·96
·4375	.....	.....	.....	.....	·1225
			·5000	·2500	·3250
1·0000	.....	.....	·5000	·2500	·3250
				·7500	·6000
·5000	.....	.....	.....	·1250	·16250
					·8375
·6500	.....	.....	.....	.....	·21125
				·62500	·43750
1·00000	.....	.....	.....	·62500	·43750
					·62625
·70000	.....	.....	.....	.....	·30625
					·32000

$$R^2 = 1{\cdot}00000 - {\cdot}32000 = {\cdot}68000$$
$$R = \sqrt{{\cdot}68000} = {\cdot}82462$$

B. *Backward Solution* — Regressions

$$\left.\begin{matrix} -{\cdot}60 \times (-{\cdot}10) \\ = {\cdot}06 \end{matrix}\right\} \left.\begin{matrix} -{\cdot}50 \times {\cdot}30 \\ = -{\cdot}15 \end{matrix}\right\} \left.\begin{matrix} -{\cdot}30 \times {\cdot}70 \\ = -{\cdot}21 \end{matrix}\right\} + {\cdot}20 = -{\cdot}10$$

$$\left.\begin{matrix} -{\cdot}625 \times {\cdot}30 \\ = -{\cdot}1875 \end{matrix}\right\} \left.\begin{matrix} -{\cdot}50 \times {\cdot}70 \\ = -{\cdot}3500 \end{matrix}\right\} + {\cdot}4375 = -{\cdot}10$$

$$\left.\begin{matrix} -{\cdot}50 \times {\cdot}70 \\ = -{\cdot}3500 \end{matrix}\right\} + {\cdot}6500 = {\cdot}30$$

$${\cdot}7000 = {\cdot}70$$

The working of the forward solution will be obvious from the figures themselves. For the backward solution the simplest procedure is to copy down in rows the column of multipliers, omitting the figure of 1·00. All the figures, except the last in each row,

should have a minus sign prefixed. The substitutions and additions are then worked from the bottom line upwards. As each regression is computed, the value found is written down against the appropriate multiplier in the rows above. If any doubt arises, the student can either refer back to the corresponding equations in Section II above or turn to the full working instructions set out in the Appendix (pp. 50–54).

IV. THE DOOLITTLE METHOD

Special Advantages. For the computer who has repeatedly to solve a number of equations of the type here described, the so-called Doolittle method [1] would seem to be the most convenient. It is widely recommended by American writers ; but, strange to say, it is seldom employed by psychologists in this country. However, after a little practice with the procedure I have described above, the student will have no difficulty in converting it into the Doolittle.

Table II shows how the ' forward solution ' would be carried out with that method. The table is self-explanatory. For further details see the working instructions given in almost any text-book, e.g., Guilford, *Psychometric Methods*, p. 395. As will be perceived on comparing it with

TABLE II. DOOLITTLE METHOD FOR SOLVING NORMAL EQUATIONS FOR REGRESSIONS

Forward Solution

Column Number ..		1	2	3	4	5
Row	Instruction					
A	r_{1k}	1·00	·60	·50	·30	·20
B	$A \div (-A1)$	−1·00	− ·60	− ·50	− ·30	−·20
C	r_{2k}		1·00	·70	·50	·40
D	$A \times B2$		− ·36	− ·30	− ·18	−·12
E	$C+D$		·64	·40	·32	·28
F	$E \div (-E2)$		1·00	·625	− ·50	−·4375
G	r_{3k}			1·00	·60	·60
H	$A \times B3$			− ·25	− ·15	−·10
I	$E \times F3$			− ·25	− ·20	−·175
J	$G+H+I$			·50	·25	·325
K	$J \div (-J3)$			− 1·00	− ·50	−·65
L	r_{4k}				1·00	·80
M	$A \times B4$				− ·09	−·06
N	$E \times F4$				− ·16	−·14
O	$J \times K4$				− ·125	−·1625
P	$L+M+N+O$				·625	·4375
Q	$P \div (-P4)$				−1·000	−·7000

Backward Solution

$$\beta_{54} = -Q5 = \cdot 7000$$
$$\beta_{53} = -K5 + \beta_{54}K4 = \cdot 3000$$
$$\beta_{52} = -F5 + \beta_{54}F4 + \beta_{53}F3 = -\cdot 1000$$
$$\beta_{51} = -B5 + \beta_{54}B4 + \beta_{53}B3 + \beta_{52}B2 = -\cdot 1000$$

[1] Doolittle, M. H., ' Methods Employed in the Solution of Normal Equations,' *U.S. Coast and Geodetic Survey Report*, 1878, pp. 115–20. The method was based on the continental versions of Gauss's procedure.

Table I, it differs from the former solely by the omission or the rearrangement of certain rows. The chief alteration is that hierarchies of the same order are collected together and subtracted all at once. As a result two improvements are made : there are (i) fewer figures to write down and (ii) fewer subtractions to carry out. Indeed, the abridgement could be carried still further, since (as will be seen) certain lines are repeated unnecessarily.

V. INVERSES

Factorial Procedure. To find the inverse (or reciprocal) of a symmetric matrix (e.g., R^{-1}, where $R = FF'$ is a correlation matrix with unity in the diagonal) we may proceed much as before : only now, on the right-hand side of our initial equation (*ib*), instead of a single column, consisting of empirical criterion correlations, we shall have n columns, consisting of $\{1, 0, \ldots, 0\}$, $\{0, 1, \ldots, 0\}, \ldots, \{0, 0, \ldots, 1\}$. In this case, too, it will be a little more expeditious to work with the standardized factor saturations. This means adopting the ' irrational ' procedure, since we now divide the top row of the residuals by the square root of the leading term. The rest of the computations will be clear from Table III.[1]

The final result of the ' forward solution ' will be twofold : we shall have (i) a triangular factor matrix (F) and (ii) the inverse of that triangular factor matrix (F^{-1}). To obtain the inverse of the correlation matrix, we have merely to multiply F^{-1} by its transpose. The matrix products FF' and $F^{-1}F'^{-1}$ are shown at the foot of Table III. FF' proves useful as a check on the working : the beginner should also calculate FF^{-1} (or $F'^{-1}F' = I$.

VI. TRIANGULAR GROUP FACTOR METHODS

Triangular Factor Matrices. In view of the basic part played by triangular matrices in the theory of linear equations and in the reduction of matrices to canonical forms,[2] it is not surprising that the triangular method of factorizing a correlation matrix should be one of the earliest forms to appear.[3] As we have seen, Yule, in his

[1] Here again we are really dealing with generalizations of procedures which are thoroughly familiar in their scalar form from elementary algebra. For example, the matrix product FF' used to be called the square of F : and what I have called the factorial method of finding an inverse is in effect an extension to matrix problems of an elementary device for finding reciprocals of scalar squares : (e.g., 625 can be factorized and expressed as 25×25 ; and since the reciprocal of 25 can be much more readily obtained than the reciprocal of 625, we take $1/625 = (1/25)^2 = (\cdot 04)^2 = \cdot 0016$).

Another familiar device can also be generalized to matrix form and may be used with advantage to obtain the inverses of large matrices. The reciprocal of $(1 - d)$, i.e., $1/(1 - d) = 1 + d + d^2 + \ldots + d^n/(1 - d)$: (see *College Algebra*, p. 163). Hence, as a first approximation to $1/(1 - d) = x^{-1}$, say, we may take $a_1 = 1 + d$. An improved approximation can now be found by taking $a_2 = a_1(2 - a_1 x) = (1 + d)(2 - 1 + d^2) = 1 + d + d^2 + d^3$: (e.g., $1/\cdot 95 = 1\cdot 00 + \cdot 05 + \cdot 0025 + \cdot 000125 = 1\cdot 052625$ approximately ; the correct value is $1\cdot 052623 \ldots$). Similarly, in matrix work, if we write $X = I - D$, we may take, as a first approximation to X^{-1}, $A_1 = I + D$. Then as an improved approximation we can take $A_2 = A_1(2I - A_1 X)$; and, if we wish, repeat the process : cf. Frazer, Duncan, and Collar, *loc. cit.*, Section 4.11, pp. 120–1 ; see also, for this and other approximations, the memorandum already cited, Section IV.

[2] MacDuffee, for example, opens his monograph on *Vectors and Matrices* (1943) by stating that any " system of homogeneous equations is equivalent to a triangular system " (theorem 2, p. 9) ; and this equivalence is basic to his subsequent proofs. Cullis, as we have seen, says that a particular solution of $XX' = A$ is ' best obtained by the *triangular* transformation ' described above (p. 356). Frazer, Duncan, and Collar, in *Elementary Matrices* (1938), start their chapter on ' numerical methods ' with a detailed demonstration of this triangular mode of reduction (p. 97 : a worked example for a symmetric matrix will be found on p. 32).

[3] As I have pointed out in an earlier discussion (*Marks of Examiners*, 1936, pp. 307f.) two important types of triangular matrices appear in factorial work. (i) A *bipolar* triangular matrix, having the form of an ' equilibrated orthogonal matrix,' emerges when we factorize a correlation matrix with disproportionately large self-correlations. The clearest instance occurs when we attempt to factorize a unit matrix : (cf. this *Journal*, I, p. 8, II, p. 112, and footnote [3] above, p. 33). (ii) A *positive* triangular matrix emerges when we take each test in turn as the basis for an orthogonal factor. This is the form with which we are here concerned.

TABLE III. THE CALCULATION OF AN INVERSE

A. *Factorization*

Correlation Matrix (R)					Unit Matrix			
Tests	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
(1)	1·00	·60	·50	·30	−1·00			
(2)		1·00	·70	·50		−1·00		
(3)			1·00	·60			−1·00	
(4)				1·00				−1·00
Multipliers	1·00	·60	·50	·30	−1·00			
1·00	1·00	·60	·50	·30	−1·00			
·60		·36	·30	·18	− ·60			
·50			·25	·15	− ·50			
·30				·09	− ·30			
		·64	·40	·32	·60	−1·00		
1st Residuals			·75	·45	·50		−1·00	
				·91	·30			−1·00
Multipliers		·80	·50	·40	·75	−1·250		
·80		·64	·40	·32	·600	−1·000		
·50			·25	·20	·375	− ·625		
·40				·16	·300	− ·500		
2nd Residuals			·50	·25	·125	·625	−1·000	
				·75	·000	·500		−1·000
Multipliers			·707	·354	·1768	·8839	−1·4142	
·707			·500	·250	·1250	·6250	−1·0000	
·354				·125	·0625	·3125	− ·5000	
3rd Residuals				·625	− ·0625	·1875	·50000	−1·0000
Multipliers				·791	− ·0791	·2372	·6325	−1·2649
·791				·625	− ·0625	·1875	·5000	−1·000

Factor Matrix (F')					Inverse of Factor Matrix (F'^{-1})			
Factor	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
1	1·00	·60	·50	·30	1·0000	·0000	·0000	·0000
2·1	·00	·80	·50	·40	− ·7500	1·2500	·0000	·0000
3·12	·00	·00	·707	·354	− ·1768	− ·8839	1·4142	·0000
4·123	·00	·00	·000	·791	+ ·0791	− ·2372	− ·6325	1·2649

B. *Product Sums*

Test	Check ($FF' = R$)				Inverse of Correlation Matrix ($F^{-1}F'^{-1} = R^{-1}$)			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
(1)	1·00	·60	·50	·30	1·60	− ·80	− ·30	·10
(2)	·60	1·00	·70	·50	− ·80	2·40	−1·10	− ·30
(3)	·50	·70	1·00	·60	− ·30	−1·10	2·40	− ·80
(4)	·30	·50	·60	1·00	·10	− ·30	− ·80	1·60

papers on partial correlation,[1] shows how a normally distributed set of n correlated variables, $x_1, x_2, x_3, \ldots, x_n$, may be expressed in terms of n uncorrelated variables, by taking successive sets of residuals, such as $x_1, x_{2\cdot1}, x_{3\cdot12}, \ldots, x_{n\cdot12\ldots(n-1)}$, where (in his notation) $x_{k\cdot12\ldots j}$ denotes the 'residual of x_i of order j.' The saturations can be determined by progressive applications of the formula for semi-partial correlation,[2] viz. :

$$r_{ig_k} = \frac{r_{ik} - \sum_j r_{ig_j} r_{kg_j}}{\sqrt{1 - \sum r^2_{kg_j}}}, \qquad j = 1, 2, \ldots, (k-1)$$

where g_k denotes the residual obtained from the kth test after tests $1, 2, \ldots, j$ have been progressively partialled out, and r_{ig_k} denotes the saturation of any test (i, say) with that residual. Thus, with the matrix of correlations given above, Yule's method of residual analysis leads at once to the standardized factor saturations (F') shown at the foot of Table III. For example, for test 4 and factor III, the formula gives

$$r_{4(3\cdot12)} = \frac{r_{43} - r_{4(1\cdot0)}\, r_{3(1\cdot0)} - r_{4(2\cdot1)}\, r_{3(2\cdot1)}}{\sqrt{1 - r^2_{3(1\cdot0)} - r^2_{3(2\cdot1)}}}$$
$$= \frac{\cdot60 - \cdot30 \times \cdot50 - \cdot40 \times \cdot50}{\sqrt{1\cdot00 - \cdot50^2 - \cdot50^2}} = \cdot354,$$

where $r_{4(3\cdot12)}$ denotes the correlation between test 4 (in its entirety) and test 3 (after the effects of tests 1 and 2 have been successively eliminated).

Together with the first test, the residuals so calculated form a set of n columns, the jth column covering $(n - j + 1)$ tests. The saturations will therefore form a triangular factor matrix ; and, as with other forms of group factor analysis, the values will remain unchanged, even if further tests are appended to the battery.

Application to Isoclinal Tables. Yule's method of analysis appeared to fit in very well with one mode of interpreting the structural relations between mental abilities. About that time, it may be remembered, a favourite theory with many psychologists maintained that the mind, on its cognitive aspect, was organized in a hierarchy of four main levels. Tests of processes on a higher level (reasoning, for example) must of necessity involve processes on the lower levels (association, perception, and sensation) ; but a test for a process on a lower level (e.g., sensory discrimination) need not involve any process on a higher level. Thus, if each level is represented by its own factor (or factors), we should expect to find a pyramid of superposed group factors, the lower and more elementary covering a wider range than the higher and more specialized.

Now, in my own investigations,[3] the tests were carefully chosen so as to represent each of the chief levels in this hierarchical structure ; and the correlations between them seemed at first sight to exhibit something like the gradation we should expect on this view. The table obtained by Spearman in his 1904 investigation displays still more clearly the steplike pattern which such a theory would imply : when his figures are smoothed, they increase by regular increments of almost exactly ·05, as shown in Table IV below.[4]

[1] Yule, G. U., *loc. cit. sup.* A slightly more elaborate discussion of this particular 'canonical transformation' is given by another worker in Pearson's department, McMahon, J., in a paper on 'Hyperspherical Goniometry and its Application to Correlation Theory for n Variables,' *Biometrika*, XI, 1923, pp. 194f. McMahon suggests that "the uncorrelated characteristics (thus derived) may be regarded as representing hidden *elementary capacities* on which the outward manifestations functionally depend." See also Camp, B. H., 'The Converse of Spearman's Two Factor Theorem,' *Biometrika*, XXXII, 1932, p. 422.

[2] See below, p. 49, footnote [1].

[3] 'Experimental Tests of General Intelligence,' *Brit. J. Psych.*, III, 1909, pp. 161f. : cf. pp. 98–9 for the scheme of levels. Also *Brit. Ass. Ann. Rep.*, 1910, pp. 804f.

[4] Spearman, C., 'General Intelligence Objectively Determined and Measured,' *Amer. J. Psychol.*, XV, 1904, p. 275. Instead of the observed coefficients (reprinted in this *Journal*, II, p. 163), I here give the smoothed version, which (except for the diagonal values) is Spearman's own : as he himself observes, "the rounded figures make the pattern clearer." For the same reason I have omitted the correlations for Music, because, as Spearman points out, these seem "rather out of place, perhaps as a result of their low reliability."

TABLE IV. ISOCLINAL INTER-CORRELATIONS

Test	i	ii	iii	iv	v
i	(1·00)	·50	·55	·60	·65
ii	·50	(1·00)	·60	·65	·70
iii	·55	·60	(1·00)	·70	·75
iv	·60	·65	·70	(1·00)	·80
v	·65	·70	·75	·80	(1·00)

A simple inspection of the table suggests that the factor responsible for the simplest test (test i, sensory discrimination) might be regarded as a general factor running through all the processes tested or examined ; whereas the factors responsible for the large coefficients towards the bottom right-hand corner might be considered more or less peculiar to the higher and more complex processes. In fact, at first sight, the pattern of correlations strongly suggests a scheme of progressively overlapping group factors.[1] Such a structural arrangement can be given a precise formulation if we reduce the observed variables or ' tests ' to terms of a hypothetical set of uncorrelated variables or ' factors ' of the type envisaged by Yule. The correlations between Spearman's tests and the factors derived in this way are set out in Table V.

TABLE V. SATURATIONS FOR COMPLETELY OVERLAPPING GROUP FACTORS

Test	Factor	x_1	$x_{2\cdot1}$	$x_{3\cdot12}$	$x_{4\cdot123}$	$x_{5\cdot1234}$	Square Sum	g
i	x_1	1·000	·000	·000	·000	·000	1·000	·69
ii	x_2	·500	·866	·000	·000	·000	1·000	·75
iii	x_3	·550	·375	·746	·000	·000	1·000	·81
iv	x_4	·600	·404	·293	·625	·000	1·000	·87
v	x_5	·650	·433	·308	·232	·491	1·000	·93
Square Sum ..		2·335	1·242	·737	·445	·241	5·000	3·317

It will be seen that the result of the analysis is a group factor matrix in which the factors vary in their range from the first factor, x_1, which extends over all the tests, to the last, $x_{5\cdot1234}$, which is specific to one test only, x_5. Moreover, each factor in its turn completely overlaps all those that come after it. For this reason such a series has commonly been described as a set of ' completely overlapping ' group factors.

The limitations of this mode of factor analysis are obvious. As will be seen from the last line of the table, it tends to spread out the contributions to the variance, and in consequence requires far more factors to fit the table (within a given margin of significance) than is absolutely necessary. We might seek to correct this by reducing the saturation for test i and raising the saturations for the other tests in the case of the first factor. But, as we lower self-correlation nearer and nearer to ·65, it becomes plain that we are virtually taking test x_5 instead of x_1 to represent the general factor. Thus, if, instead of assuming (with Spearman) that the first and most general factor must consist of the lowest and simplest of all the processes tested, we identify it (approximately at any rate) with the highest and most complex, we shall find it possible to account for nearly all the common factor variance with a single comprehensive factor only : (see last column, headed g, in Table V).[2] Thus the gradational mode of interpreting the ' hierarchy ' of mental structure, therefore, plausible as it may seem at first sight, here falls to the ground. The fact is that efficiency on the different mental levels is positively correlated, and so cannot be attributed to separate and independent group factors

[1] The chief advocate of the above mode of interpreting Spearman's table and my own was Professor J. A. Green (secretary of the Committee on Mental and Physical Factors) : he, however, supposed that the factor saturations would also go by regular steps—though, so far as I am aware, he never carried out any actual mathematical analysis to test whether this would be so. The factor matrix printed in the text and much of the above discussion are taken from a memorandum which I drew up for the Committee (1915).

[2] Cf. also ' The Two-Factor Theory,' this *Journal*, II, Table III, p. 163. Much the same proved to be the case with the correlations obtained in my own investigations. (*loc. cit. sup.*) : indeed, one general factor supplemented by four non-overlapping group factors for the various levels yields a much better fit.

until individual differences in the general efficiency of the mind as a whole, on all its various levels, has been fully allowed for.[1]

Other Practical Applications. In recent years this method of analysis has come to be regarded as of indirect or theoretical interest only. In practical work it has been used mainly to provide a convenient preliminary simplification [2] before computing more important values (e.g., to reduce singular matrices to non-singular matrices of a lower order, to secure orthogonal factors for further rotation, to evaluate the determinants required for tests of mathematical consistency, etc.) : and even these applications are rare. Nevertheless, in certain types of investigation, I still consider that the method may be of direct practical service. As I have pointed out elsewhere,[3] it seems particularly appropriate where a series of tests has been so selected that they represent cumulative degrees of factorial complexity.

Many batteries, for example, include a test of intelligence assumed to enter into all the other tests ; and in such a case the intelligence test can be taken to represent the first factor, and the remaining factors can be treated as group factors and extracted in order according to the decreasing range of tests over which they extend. In applying the ' triangular method ' for this purpose, it should be noted (i) that the self-correlations need not be unity, (ii) that the columns may be shortened by more than a single test at a time, (iii) that one or more of the infra-diagonal elements may be zero or approximately zero, and (iv) that the method can usefully be employed in conjunction with the ' sum method of factorization ' : (see this *Journal*, II, pp. 61–3).

The method is particularly appropriate in analysing results obtained with tests of learning or maturation, since both these processes are largely cumulative. With the former one of the earliest and clearest instances appeared in the analysis of arithmetic tests undertaken by Professor J. A. Green and myself for the Brit. Ass. Committee on ' Mental and Physical Factors in Education ' (*Ann. Rep.*, LXXXV, 1916, pp. 307f.) : a child cannot subtract until it has learnt to add, or divide until it has learnt to subtract, or work problems in fractions before it can divide ; hence each of these processes gives rise to group factors fitting into a triangular scheme. Among tests of maturation the progressive items in the Binet scale rest on a similar assumption (e.g., the successive tests of memory for words and for numbers, the Terman vocabulary tests, etc.). In such cases the saturations for any one factor usually diminish from the easier tests to the harder (i.e., they decrease in the opposite direction from that shown in Table V). Hence the resulting correlation table tends to exhibit a diagonal ridge ; and this, I suggest, may commonly be regarded as the cue for applying the triangular method.

Many of the correlation tables which have been analysed by more circuitous procedures might be conveniently dealt with by this shorter procedure. For instance, the first of the tables which Holzinger prints to exemplify his own bifactor analysis can readily be factorized in this way by taking test 7 as the first factor. Again, the rotated matrix of overlapping group factors which Thurstone gives to illustrate his method of determining ' primary traits ' can also be reached more directly by the procedure described above, without first undertaking a centroid analysis and then applying a tentative set of rotations.[4] If I am right, then many factorial studies modelled on one or other of these illustrations could also be dealt with in this fashion.

VII. CALCULATION OF DETERMINANTS

Residual Coefficients in Determinantal Form. The method of hierarchical subtraction will be found useful for computing the determinant and minors of the correlation matrix, and the various statistical constants derived from it which can be

[1] Cf. Burt, C., *loc. cit.* (1909), p. 65.

[2] If we apply Yule's triangular method to a unit matrix, we obtain a ' standard numerical form ' (as I have called it) which has many helpful uses : (cf. *Report* of Brit. Ass. Committee, *loc. cit. sup.*, also this *Journal*, I, p. 8, and II, p. 112, Table VB). With the positive factor omitted, it makes a useful ' reduction matrix ' (cf. above, p. 33). It also yields a form of analysing variance comparable with the factorial analysis described above : (for examples, see Mather, K., *Statistical Analysis in Biology*, pp. 61f.).

[3] This suggestion is discussed more fully in *The Factors of the Mind*, p. 307. A concrete example is given below, p. 49.

[4] Holzinger, K. J., *Student Manual of Factor Analysis*, 1937, Table IV, p. 12. Thurstone, L. L., *The Vectors of Mind*, 1935, Table 4, p. 169. It should be noted that, when the object is a factorial analysis, it is wrong to eliminate the first factor by using the *full* formula for partial correlation, as is done in so many research-theses. Which of the two formulæ is appropriate must depend on the question to be answered.

expressed in determinantal form. It will therefore be helpful to indicate, in formal symbols, what are the chief determinants that emerge incidentally in the course of the foregoing calculations.

Let us re-write the initial set of equations we have to solve in the following way :

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots &= a_{1m} \\ a_{21}x_1 + a_{22}x_2 + \ldots &= a_{2m} \\ \ldots \quad \ldots \quad \ldots &\quad \ldots \end{aligned}$$

Confining ourselves for the moment to the first two unknowns and the first two equations, the *first* hierarchy and residuals can readily be expressed in determinantal form as shown below.

1st Hierarchy

Multipliers				
1	$a_{11}x_1$	$+ a_{12}x_2$	$=$	a_{1m}
$\frac{a_{21}}{a_{11}}$	$\frac{a_{21}}{a_{11}}a_{11}x_1$	$+ \frac{a_{21}}{a_{11}}a_{12}x_2$	$=$	$\frac{a_{21}}{a_{11}}a_{1m}$

1st Residuals

$$\begin{aligned} 0 \quad + \quad 0 \quad &= \quad 0 \\ 0 \quad + \left(a_{22} - \frac{a_{21}}{a_{11}}a_{12}\right)x_2 &= a_{2m} - \frac{a_{21}}{a_{11}}a_{1m} \end{aligned}$$

Cleared of Fractions

$$\frac{1}{a_{11}}(a_{11}a_{22} - a_{21}a_{12})x_2 = \frac{1}{a_{11}}(a_{11}a_{2m} - a_{21}a_{1m})$$

Expressed as Determinants

$$\frac{1}{a_{11}}\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} x_2 = \frac{1}{a_{11}}\begin{vmatrix} a_{11} & a_{1m} \\ a_{21} & a_{2m} \end{vmatrix} .$$

x_1 has now been eliminated, and we have reached a determinantal equation for x_2 which is in accordance with Cramer's rule.

If we now extend the procedure to the first three unknowns and the first three equations, then, after eliminating x_2 as well as x_1, we reach [1] an analogous expression for x_3, namely,

$$\frac{1}{\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}}\begin{vmatrix} a_{11}a_{12}a_{13} \\ a_{21}a_{22}a_{23} \\ a_{31}a_{32}a_{33} \end{vmatrix} x_3 = \frac{1}{\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}}\begin{vmatrix} a_{11}a_{12}a_{1m} \\ a_{21}a_{22}a_{2m} \\ a_{31}a_{32}a_{3m} \end{vmatrix}$$

All this can be formulated much more concisely if we avail ourselves of one of the traditional conventions for denoting a determinant by a single letter. Cullis adopts a modification of the so-called ' umbral ' notation, with subscripts to indicate row numbers and (where necessary) superscripts to indicate column numbers. Thus we may use

$$d_{123}^{12n} \text{ to mean } \begin{vmatrix} a_{11}a_{12}a_{1n} \\ a_{21}a_{22}a_{2n} \\ a_{31}a_{32}a_{3n} \end{vmatrix} .$$

Where the determinant is symmetrical, the superscripts can be omitted ; d_1 of course will denote a_{11} ; and we can take $d_0 = 1$. We can now express residual equations like the above, which arise from the progressive elimination of $x_1, x_2, \ldots$, in a succinct form as follows :

$$\frac{1}{d_0}\{ d_1x_1 + d_2x_2 + d_3x_3 + \ldots \} = \frac{1}{d_0}d_1^m,$$

$$\frac{1}{d_1}\{ d_{12}x_2 + d_{13}x_3 + \ldots \} = \frac{1}{d_1}d_{12}^{1m},$$

$$\frac{1}{d_{12}}\{ d_{123}x_3 + \ldots \} = \frac{1}{d_{12}}d_{123}^{12m}, \text{ etc.}$$

[1] To assist the student in his attempt to verify this further step, it may be useful to remind him that

$$\begin{vmatrix} \begin{vmatrix} a_{11}a_{12} \\ a_{21}a_{22} \end{vmatrix} & \begin{vmatrix} a_{11}a_{1n} \\ a_{21}a_{2n} \end{vmatrix} \\ \begin{vmatrix} a_{11}a_{12} \\ a_{31}a_{32} \end{vmatrix} & \begin{vmatrix} a_{11}a_{1n} \\ a_{31}a_{3n} \end{vmatrix} \end{vmatrix} = a_{11}\begin{vmatrix} a_{11}a_{12}a_{1m} \\ a_{21}a_{22}a_{2m} \\ a_{31}a_{32}a_{3m} \end{vmatrix}$$

See, for example, Aitken, A. C., *Determinants and Matrices* (1939), p. 48.

Each of these equations corresponds with the first row of residual equations obtained on p. 35, and tabulated in Table I. Substituting the numerical values we have

$$\frac{1}{1}\{\ 1{\cdot}00x_1 + {\cdot}60x_2 + {\cdot}50x_3 + {\cdot}30x_4\ \} = \frac{1}{1}\ {\cdot}20 = {\cdot}20 \qquad (a.\text{i})$$

$$\frac{1}{1{\cdot}00}\{\ {\cdot}64x_2 + {\cdot}40x_3 + {\cdot}32x_4\ \} = \frac{1}{1{\cdot}00}\ {\cdot}28 = {\cdot}28 \qquad (c.\text{ii})$$

$$\frac{1}{{\cdot}64}\{\ {\cdot}32x_3 + {\cdot}16x_4\ \} = \frac{1}{{\cdot}64}\ {\cdot}208 = {\cdot}325 \qquad (e.\text{iii})$$

$$\frac{1}{{\cdot}32}\{\ {\cdot}20x_4\ \} = \frac{1}{{\cdot}32}\ {\cdot}14 = {\cdot}4375 \qquad (g.\text{iv})$$

The method of hierarchical subtraction does not give us the numerators and denominators separately : e.g., in Table I we do not find $\frac{1}{{\cdot}32}$ and ·20, but only their product ·625. Hence to calculate any of the larger determinants we have to effect a cancellation of the unwanted denominator by multiplication. Thus to find d_{123} we may take $\frac{1}{d_1}d_{12} \times \frac{1}{d_{12}}d_{123} = {\cdot}64 \times {\cdot}50 = {\cdot}32$ (the figure in Table I). Again, to find d_{123}^{12m} we may take $\frac{1}{d_1}d_{12} \times \frac{1}{d_{12}}d_{123}^{12m} = {\cdot}64 \times {\cdot}325 = {\cdot}208$. And to find d_{12m} we may take $\frac{1}{d_1}d_{12} \times \frac{1}{d_{12}}d_{12m} = {\cdot}64 \times {\cdot}8375 = {\cdot}536$.

The Determinant of the Correlation Matrix. It follows that the product of the coefficients in the leading terms of the residual matrices, which are successively obtained by hierarchical subtraction, will give the determinant for the entire matrix : for $\frac{1}{d_0}d_1 \times \frac{1}{d_1}d_{12} \times \frac{1}{d_{12}}d_{123} \times \frac{1}{d_{123}}d_{1234} = d_{1234} = \Delta$ (say). Thus, taking figures from the equations above, we have $\Delta = 1{\cdot}00 \times {\cdot}64 \times \frac{{\cdot}32}{{\cdot}64} \times \frac{{\cdot}20}{{\cdot}32}$, or, taking the figures from Table I, $= {\cdot}64 \times {\cdot}50 \times {\cdot}625 = {\cdot}200$. The procedure thus yields a simple device for calculating large determinants.

Factor Saturations in Determinantal Form. Remembering that the equation $R = FF'$ is a matrix generalization of the scalar equation $r = f \times f$, i.e., f^2, we may regard the extraction of a factor matrix F from a symmetrical matrix R as the analogue of the extraction of a square root. A little algebra will show that the factor saturations we require will be obtained from the triangular equations set out above by dividing the terms in the brackets by two square roots. In determinantal form, therefore, they can be expressed as follows : (the divisors, $\sqrt{d_0}$ and $\sqrt{d_1}$, in the first line are inserted for consistency ; the reader will remember that $d_0 = 1$ and that with self-correlations of unity d_1 also $= 1$).

The saturations will be :

$$\text{I} \quad \frac{1}{\sqrt{d_0}\sqrt{d_1}}\ \{\ d_1, \qquad d_2, \qquad d_3\ , \ldots\ \}$$

$$\text{II} \quad \frac{1}{\sqrt{d_1}\sqrt{d_{12}}}\ \{\ d_{12}, \qquad d_{12}^{13}\ , \ldots\ \}$$

$$\text{III} \quad \frac{1}{\sqrt{d_{12}}\sqrt{d_{123}}}\ \{\ d_{123}, \ldots\ \}$$

and generally [1]

$$\frac{1}{\sqrt{d_{12\ldots(n-1)}}\sqrt{d_{12\ldots n}}}\ \{d_{12\ldots n}, \qquad d_{12\ldots n}^{12\ldots(n+1)}, \qquad \ldots\ \}$$

The validity of these expressions can readily be tested by referring once again to Table I. We have

$$\text{I} \quad 1{\cdot}00, \qquad {\cdot}60, \qquad {\cdot}50, \qquad {\cdot}30$$

$$\text{II} \quad \frac{1}{\sqrt{1{\cdot}00}\sqrt{{\cdot}64}}\ \{\ {\cdot}64, \qquad {\cdot}40, \qquad {\cdot}32\ \}$$

$$= {\cdot}80, \qquad {\cdot}50, \qquad {\cdot}40$$

$$\text{III} \quad \frac{1}{\sqrt{{\cdot}64}\sqrt{{\cdot}32}}\ \{\ {\cdot}32, \qquad {\cdot}16\ \}$$

$$= {\cdot}707, \qquad {\cdot}354$$

$$\text{IV} \quad \frac{1}{\sqrt{{\cdot}32}\sqrt{{\cdot}20}}\ \{\ {\cdot}20\ \}$$

$$= {\cdot}791$$

[1] Except for differences in the notation this expression is identical with that which Cullis gives as " the best particular solution " of the equation $XX' = A$: *loc. cit.*, II, pp. 356–8, Theorem 1, ex. viii, case I ; also p. 282, ex. iv.

In each case the values given by the determinantal formula reduce to those found by the method of hierarchical subtraction.

Criterion for Consistency. If (as ought to be the case with any set of inter-correlations which have been correctly calculated) the initial correlation matrix is ' positive definite ' (or ' semi-definite '), then all the discriminants of whatever order must be positive.[1] If not, it will be impossible to factorize the table in terms of ' real ' components (i.e., ' factors ' whose saturations do not involve imaginary numbers). We thus have a convenient criterion for determining whether or not the values calculated for a particular set of inter-correlations are mutually consistent. With an empirical table of correlations, *all the leading principal minors should be non-negative.*[2]

By way of an example, let us consider once again the kind of table originally put forward by Spearman to illustrate his conception of a hierarchy. In explaining the general principle, his ' ideal table ' usually included ' reliability coefficients ' reduced to fit the regular step-like progression. Thus his original version of the perfect hierarchy was as follows.

TABLE VI. ISOCLINAL MATRIX WITH REDUCED SELF-CORRELATIONS

Test	i	ii	iii	iv
i	·45	·50	·55	·60 . . .
ii	·50	·55	·60	·65 . . .
iii	·55	·60	·65	·70 . . .
iv	·60	·65	·70	·75 . . .

Here, as will be seen at once, the leading minors are negative, e.g.,

$$\begin{vmatrix} \cdot 45 & \cdot 50 \\ \cdot 50 & \cdot 55 \end{vmatrix} = \cdot 2475 - \cdot 2500 = -\cdot 0025\ ; \quad \begin{vmatrix} \cdot 45 & \cdot 60 \\ \cdot 60 & \cdot 75 \end{vmatrix} = \cdot 3375 - \cdot 3600 = -\cdot 0225, \text{ etc.}$$

Hence in practice it would be quite impossible to obtain correlations like the above from actual data.[3]

VIII. PARTIAL CORRELATIONS

Determinantal Formula. The various figures obtained by the foregoing procedure enable coefficients of partial correlation, of any order, to be obtained quite readily. The mode of calculation can best be explained by reference to the same example as before.

Suppose we wish to calculate the partial correlation between test 3 and the criterion z, with tests 1 and 2 partialled out. Let Δ denote the determinant of the matrix containing the

[1] Suppose that a quadratic form $x'Ax$ has been reduced as above by a congruent transformation to its normal or canonical shape, $y'Vy$, where V is diagonal : then, if all the elements in V are of the same sign, the form is said to be ' definite ' (or, if some are zero, ' semi-definite ') ; and if they are all of positive sign, it is said to be ' positive.' (Cf. Cullis, II, pp. 301–3, Frazer, *et al.*, pp. 30–3.)

The ' discriminant of the form ' is the determinant $|A|$. If one or more of the variables x are omitted, we shall obtain discriminants of the forms so derived ; and these discriminants will be of a lower order. As we have seen, they are closely related to the determinantal factors from which the main discriminant can be obtained.

The criterion reached above may be regarded as a generalization of the use made of the sign of the ' discriminant,' $b^2 - 4ac$, in determining whether the roots of an ordinary quadratic equation are real and distinct (see *College Algebra*, p. 304) : cf. above, p. 37.

[2] An instructive instance of this problem has been raised by Mr. Lubin and others who have criticized the correlation table published by Sheldon (see Lubin, A., this *Journal*, III, pp. 187–9). In such a case I myself should employ the criterion proposed in the text : we can take the maximum value for the self-correlations, namely, 1·00, insert it throughout the leading diagonal, and then factorize the table by the triangular method already described ; it thereupon becomes quite clear that Sheldon's correlations as printed are mutually inconsistent.

[3] Cf. this *Journal*, III, p. 169 and refs. Spearman's figures were admittedly artificial, and intended only to explain how his 1904 results showed figures " diminishing in each row by approximately equal amounts." But the above criticism, due in the first instance to Sheppard, helped to persuade him to substitute the synclinal or ' equi-proportional ' form of hierarchy for the isoclinal or ' additive.'

relevant correlations, i.e., in the present case, the matrix shown in the top panel of Table I with data for test 4 deleted (since that test does not enter into the problem); and let Δ_{ij} denote a minor in this determinant obtained by striking out the ith row and the jth column. Then Pearson's determinantal formula for the partial correlation [1] may be written

$$r_{3z\cdot 12} = \frac{\Delta_{3z}}{\sqrt{\Delta_{33}}\sqrt{\Delta_{zz}}}, \qquad \text{(vii)}$$

or, in the notation used above,

$$= \frac{d^{12z}_{123}}{\sqrt{d_{12z}}\sqrt{d_{123}}}. \qquad \text{(vii}a\text{)}$$

On referring to the values obtained by hierarchical subtraction, in the preceding section, we have

$$\begin{aligned} d^{12z}_{123} &= \cdot 64 \times \cdot 325 = \cdot 208, \\ d_{12z} &= \cdot 64 \times \cdot 8375 = \cdot 536, \\ d_{123} &= \cdot 64 \times \cdot 50 = \cdot 32. \end{aligned}$$

The leading principal minor which forms the first factor (·64) is common to all three determinants, and may therefore be cancelled out. We can accordingly calculate the partial correlation direct from three figures in the residual matrix forming the third panel in Table I. We thus obtain :

$$r_{3z\cdot 12} = \frac{\cdot 325}{\sqrt{\cdot 8375}\sqrt{\cdot 50}} = \cdot 5023. \qquad \text{(viii)}$$

Or again, suppose we desire to calculate the partial correlation between two consecutive tests, say, tests 3 and 4, with the preceding tests partialled out : our formula gives

$$r_{34\cdot 12} = \frac{d^{124}_{123}}{\sqrt{d_{123}}\sqrt{d_{124}}} = \frac{\cdot 25}{\sqrt{\cdot 50}\sqrt{\cdot 75}} = \cdot 408. \qquad \text{(ix)}$$

The general rule will now be clear. The procedure, we observe, is somewhat similar to that adopted in correcting for unreliability. Following Yule let us call the two tests or variables in whose correlation we are interested the ' primary or wanted ' variables, and those which are to be partialled out the ' secondary or unwanted ' variables. Select the panel which gives the residual correlations remaining after the ' unwanted ' variables have been eliminated : then take as numerator the residual correlation between the two ' wanted ' variables, and as denominator the geometric mean of the residual self-correlations (or ' reliabilities ') of the ' unwanted ' variables.

Semi-partial Correlation. To answer certain types of problem it is sometimes necessary to partial out an irrelevant or ' unwanted ' influence from *only one* of the two primary variables, not from both. This can be effected by Yule's formula for semi-partial correlation.[2] The formula is obtained, as we might expect, by dropping one of the two residual self-correlations from the denominator, namely, that relating to the primary variable which is *not* to be ' corrected.' Thus, if we wish to remove the influence of tests 1 and 2 from test 3 but not from test 4, we omit the determinantal divisor $\sqrt{d_{124}}$ from the above equation (ix). The ' semi-partial correlation ' will therefore be

$$\frac{d^{124}_{123}}{\sqrt{d_{123}}} = \frac{\cdot 25}{\sqrt{\cdot 50}} = \cdot 354$$

On referring back to the construction of the hierarchies in Table I (p. 35, eq. (*f*.iv), marginal value), we see that this expression and this figure represent the *saturation coefficient*

[1] Kelley, T. L., *Statistical Method*, 1923, p. 298, eq. 267.

[2] Cf. Guilford, J. P., *Psychometric Methods*, 1936, p. 404, eq. 187. A fuller discussion will be found in Dunlap, J. W., and Cureton, E. E., ' On the Analysis of Causation,' *J. Educ. Psych.*, XXI, 1930, pp. 663–72.

for test 4 with factor III. Hence, with the mode of analysis we have here been considering, *the factor saturations are semi-partial correlations.*[1]

Example. The procedure may be illustrated from what was, I believe, one of the earliest attempts to apply the method of partial correlation to the results of mental tests.[2] With a group of 318 pupils aged 7 to 14, assessments were obtained for (1) chronological age (in years), (2) intelligence (in mental years), (3) school attainments, (4) mental age with the original Binet-Simon scale. The inter-correlations are shown in Table VIIA.

From our antecedent knowledge of the variables concerned, it seems clear that we can assume that (i) chronological age (*A*, say) depends on no other variable ; (ii) intelligence as measured by mental age must depend partly on chronological age, partly on a factor of innate ability (*I*, say), but not on school attainments (unless the method of assessment is faulty) ; (iii) school attainments must depend partly on age, partly on innate ability, and partly on a third factor, which we can call scholastic efficiency (*S*, a composite factor depending mainly on the interest and diligence of the child, sometimes on his memory and verbal facility, and occasionally no doubt on the efficacy of the teaching) ; (iv) finally, the results obtained with the Binet tests (*B*), which in their original form included tests of reading, writing, mental arithmetic, etc., must undoubtedly depend on all three, as well as on possible factors of their own. On factorizing Table VIIA in accordance with these assumptions, we obtain the saturations shown in Table VIIB.

TABLE VII. FACTORIZATION OF BINET TESTS

Variable	A. Inter-correlations				B. Factor Saturation			
	1	2	3	4	1·0	2·1	3·12	4·123
1. Age (*A*)	(1·00)	·70	·87	·83	1·0000	·000	·000	·000
2. Intelligence (*I*) ..	·70	(1·00)	·75	·84	·700	·714	·000	·000
3. School Att. (*S*) ..	·87	·75	(1·00)	·91	·870	·197	·453	·000
4. Binet (*B*)	·83	·84	·91	(1·00)	·830	·363	·256	·107

From this factorization we can at once obtain the partial regressions by the method described above. We find

$$B = \cdot 56S + \cdot 34I + \cdot 11A \, .$$

And from the determinants emerging in the course of the factorization we can obtain the various partial correlations : e.g.,

$$(a)\ r_{32\cdot 1} = \frac{d^{13}_{12}}{\sqrt{d_{12}}\sqrt{d_{13}}} = \frac{\cdot 141}{\sqrt{\cdot 51}\sqrt{\cdot 243}} = \cdot 40$$

$$(b)\ r_{43\cdot 12} = \frac{d^{124}_{123}}{\sqrt{d_{123}}\sqrt{d_{124}}} = \frac{\cdot 228}{\sqrt{\cdot 400}\sqrt{\cdot 352}} = \cdot 61 \, .$$

We thus discover (*a*) that, when differences in chronological age have been eliminated, the assessments for school attainments show an appreciable correlation with the assessments for innate ability ; and (*b*) that, when both age and innate ability have been eliminated, there still remains a fairly large correlation between scholastic efficiency and performance in the Binet-Simon tests.[3]

[1] The difference between a partial and a semi-partial correlation can best be explained by taking the case of three variables only. Assuming all variables to be in standard measure, we may say that the partial correlation $r_{32\cdot 1} = \frac{r_{32} - r_{31}r_{21}}{\sqrt{1 - r^2_{31}}\sqrt{1 - r^2_{21}}}$ measures the correlation between $x_{3\cdot 1} = x_3 - r_{31}x_1$ and $x_{2\cdot 1} = x_2 - r_{21}x_1$, where the subscripts indicate that the influence of x_1 has been eliminated from both variables ; the semi-partial correlation $r_{3(2\cdot 1)} = \frac{r_{32} - r_{31}r_{21}}{\sqrt{1 - r^2_{21}}}$ is the correlation between x_3 and $x_{2\cdot 1}$, where the influence of x_1 has been eliminated from x_2 only. As we have seen, with the present mode of factor analysis, $x_{2\cdot 1}$ is identical with factor II.

[2] *Mental and Scholastic Tests*, 1921 ed., pp. 180–3 (1947 ed., pp. 192–5). The reader may find it instructive to use the above table as an exercise, and to calculate for himself the determinantal residuals, the factor saturations, the partial regressions, and the partial correlations, by the methods described in the text. This earlier analysis is still quoted to show how wrong it is to deduce the fact of causal influence merely from the discovery of a significant partial regression (e.g., Goodenough, F., *Mental Testing*, 1950, p. 253). I can find nothing in my argument to suggest such a fallacy : whether (supposing a numerical relation to be confirmed) x is a cause, an effect, or a concomitant of y is a supposition that depends, as I have repeatedly insisted, on " *antecedent* sources of knowledge " (cf. *loc. cit.*, 1947 ed., p. 279). However, the persistence of such criticisms serve to emphasize how important it is for the statistical psychologist to state explicitly precisely what his assumptions are.

[3] The complete series of partial correlations will be found in Table XX, *loc. cit.*, p. 194.

IX. SUMMARY

1. A general factorial method has been put forward for the solution of simultaneous linear equations. In the form of successive ' hierarchical subtractions ' this yields a simple working method for practical computations.

2. The method has two forms, a rational and an irrational. The former appears most suitable for such problems as the calculation of partial regressions and multiple correlations ; the latter for the calculation of inverses, determinants, partial correlations, and analysis into ' completely overlapping ' group factors.

3. The underlying principles are treated as generalizations of elementary algebraic processes. The practised student will find it easy, with minor abridgements and re-arrangements, to convert the method here described into the better known methods used by experienced computers, e.g., the Doolittle.

APPENDIX

THE CALCULATION OF PARTIAL REGRESSIONS AND MULTIPLE CORRELATIONS BY HIERARCHICAL SUBTRACTION

Working Procedure

The following method [1] of calculating partial regression coefficients and multiple correlations is recommended for those who are undertaking such calculations for the first time and are not likely to make frequent use of the procedure. By omitting certain sections of the table, or altering the arrangement of columns and rows, it can readily be converted into one or other of the many variants of the traditional procedure, e.g., that of Whittaker and Robinson, Doolittle, or Aitken. To enable comparisons to be made with other modes of working, the example chosen for illustration is drawn from Alexander's monograph on *Intelligence, Concrete and Abstract* (1935). The figures are those used by Godfrey Thomson [2] in his *Factorial Analysis of Human Ability* (1939), p. 91.

The following are the essential steps.

I. *Determining the Order in which the Tests are Taken.*

1. Arrange the tests in the order in which it is anticipated that they will contribute to the estimation of the criterion, i.e., in the order of the probable weight which each will receive in the final regression equation. Other things being equal the test obtaining the largest weight will be the test with (*a*) the largest criterion correlation, and (*b*) the smallest inter-test correlations. As a rough guide, take the ratio $\frac{\text{criterion correlation}}{\text{sum of test correlations}}$ (the ' test correlations ' to include the self-correlations

[1] The working instructions are reproduced from the memorandum on ' The Solution of Simultaneous Linear Equations ' which I prepared for the War Office Directorate of Selection of Personnel. In all three fighting services, researches on selection tests turned largely on the calculation of multiple correlations and partial regressions ; and hence a simple routine scheme became desirable at an early stage of the work. Professor Vernon, in describing the method, adds that " it is one of the quickest and easiest to learn " (*Notes on Statistical Methods*, 1945, ' Multiple Correlation,' p. 6).

[2] The figures are not precisely those given by Alexander, but have been slightly modified by Thomson. Thomson takes the tests in the order in which they are described in Alexander's monograph ; and thus, with the validity coefficients that he uses, leaves the best till last. The disadvantage of this will be seen when we consider the statistical significance of the regressions eventually found (see below, p. 54). As we shall see, the Analogies Test is certainly non-significant, and would have to be omitted ; and having reached this conclusion we should have to work the calculations afresh, practically from the start. If, however, the right order can be selected from the outset, the omission of the later tests will make no difference to the figures obtained for the earlier. It should be noted that, in calculating factor-measurements from a larger correlation matrix, which has already been factorized in terms of r common and n specific factors ($R = F_c F_c' + F_s^2$), a good approximation can be reached by taking $F_c' R^{-1} = (F_c' F_s^{-2} F_c + 1)^{-1} F_c' F_s^{-2}$: this is equivalent to computing the inverse of a smaller $r \times r$ matrix instead of the larger $n \times n$ matrix. After the matrix in brackets has been computed, the working procedure will be that shown in Table VIII.

taken as unity). With the table which here serves as an example (Table VIII, below), the ratios so obtained are as follows : (I give the tests first in the order adopted by Thomson).

(*b*) Stanford Binet Test .. .. .. 28·0 per cent.
(*c*) Thorndike Reading Test .. .. 25·7 per cent.
(*d*) Analogies Test .. .. 19·2 per cent.
(*a*) Picture Completion Test .. .. 34·1 per cent.

In accordance with these figures, therefore, the Picture Test should be put first, and the remainder in the order indicated by the letters.

2. Prepare a working sheet on the lines indicated in Table VIII. With four tests we shall need $(2n + 1) = 9$ panels with the number of rows in each decreasing as shown. In the top panel copy out the matrix of inter-test correlations in the re-arranged order, with 1·000 in the leading diagonal, and border it with the criterion correlations : i.e., write the latter in a row beneath and in a column to the right (labelled *x*), to preserve the symmetry. As a check on the working, calculate the total of each row, and enter it in the column after the criterion correlations.

TABLE VIII. CALCULATION OF MULTIPLE CORRELATION AND REGRESSIONS

A. *Forward Solution*

Test		*a*	*b*	*c*	*d*	*x*	Check Total	Squares	Ratio
	a	1·0000	·3900	·1900	·2700	·6300	2·4800	(·3969)	(·3969)
	b	(·3900)	1·0000	·6900	·4900	·7200	3·2900	(·5184)	(·5184)
	c	(·1900)	(·6900)	1·0000	·3800	·5800	2·8400	(·3364)	(·3364)
	d	(·2700)	(·4900)	(·3800)	1·0000	·4100	2·5500	(·1681)	(·1681)
	x	(·6300)	(·7200)	(·5800)	(·4100)	1·0000	3·3400		
a	1·0000	1·0000	·3900	·1900	·2700	·6300	2·4800		
b	·3900		·1521	·0741	·1053	·2457	·9672		
c	·1900			·0361	·0513	·1197	·4712		
d	·2700				·0729	·1701	·6696		
x	·6300					·3969	1·5624		
b			·8479	·6159	·3847	·4743	2·3228	·2250	·2653
c				·9639	·3287	·4603	2·3688	·2119	·2198
d					·9271	·2399	1·8804	·0576	·0621
x						·6031	1·7776		
b	1·0000		·8479	·6159	·3847	·4743	2·3228		
c	·7264			·4474	·2794	·3445	1·6872		
d	·4537				·1745	·2152	1·0538		
x	·5594					·2653	1·2993		
c				·5165	·0493	·1158	·6816	·0134	·0260
d					·7526	·0247	·8266	·0006	·0008
x						·3378	·4783		
c	1·0000			·5165	·0493	·1158	·6816		
d	·0954				·0047	·0110	·0650		
x	·2242					·0260	·1528		
d					·7479	·0137	·7616	·0002	·0003
x						·3118	·3255		
d	1·0000				·7479	·0137	·7616		
x	·0183					·0003	·0140		
						·3115	·3115		

B. *Backward Solution*

	(*a*)	(*b*)	(*c*)	(*d*)	Regressions
(*a*)	$-\cdot3895 \times \cdot3900$ $-\cdot1519$	$-\cdot2225 \times \cdot1900$ $-\cdot0423$	$-\cdot0183 \times \cdot2700$ $-\cdot0050$	$+\cdot6300$	$=\cdot4308$
(*b*)		$-\cdot2225 \times \cdot7264$ $-\cdot1616$	$-\cdot0183 \times \cdot4537$ $-\cdot0083$	$+\cdot5594$	$=\cdot3895$
(*c*)			$-\cdot0183 \times \cdot0954$ $-\cdot0017$	$+\cdot2242$	$=\cdot2225$
(*d*)				$\cdot0183$	$=\cdot0183$

C. *Check on Regressions*

Regressions	Test Correlations (*d*)	Product	Criterion Correlations (*x*)	Product
·4308	·27	·1163	·63	·2714
·3895	·49	·1909	·72	·2805
·2225	·38	·0845	·58	·1291
·0183	1·00	·0183	·41	·0075
		·4100 $= r_{dx}$		·6885 $= R^2$

D. *Multiple Correlations*

Test	*a*	*b*	*c*	*d*	*x*	Total
Variances	·3969	·2653	·0260	·0003	·3115	1·0000
Cumulative Sum	·3969	·6622	·6882	·6885	1·0000	
Square Root	·6300	·8138	·8295	·8298		
	r_{xa}	$r_{x \cdot ab}$	$r_{x \cdot abc}$	$r_{x \cdot abcd}$		

II. *Forward Solution.*

The forward solution consists in a repetition of two alternate processes, similar to those involved in factorizing a correlation matrix by simple summation : (i) first, a hierarchy is calculated and set down below the given correlation matrix (or the residuals) ; (ii) then the calculated hierarchy is subtracted from these correlations (or the residuals). And the double process is repeated until all the residuals are zero. For routine calculations the student will find it easier to leave a blank line beneath each row of correlations (or residuals), as he copies them, and then enter the corresponding row of the hierarchy, as he calculates it, immediately beneath the figures from which it has to be subtracted. Here, however, to make the several steps quite clear, I have printed each hierarchy in a separate panel below the panel for the correlations (or residuals).

3. Adopting this latter arrangement, re-write, in the first row of the second panel, the top line of the correlation table just above it. In the margin to the left, copy again the same five figures in a vertical column. Now use this column and row to construct a hierarchy by multiplying in turn each of the figures at the top by each of the figures at the side.

The sum in the check column of the top panel (2·4800) should also be multiplied at the same time. In each row the sum of the new figures entered in the second panel should be equal to the product so obtained from the check-total. The beginner will find it convenient to keep two columns for checking—one giving the product of the sum and the other the sum of the products. (To save time the experienced computer will omit the repeated figures below the diagonal of each symmetric matrix : but in checking they will have to be added in from the appropriate section of the corresponding column.)

4. Subtract this hierarchy from the matrix of observed correlations in the top panel, and write the remainders in the third panel. The sum of the remainders in each row should be equal to the remainder obtained on subtracting the corresponding sums in the check column.

5. Now copy the top row of this residual matrix to form the top row in the fourth panel. Then divide each of the residuals by the first (·8479) ; and write the quotients in a column on the left. As before multiply in turn each of the figures in the top row of the panel by each of the quotients at the side. In this way a second hierarchy will be constructed. Add each row, and compare the sums with the figure obtained in the check column.

6. Subtract this hierarchy from the figures in the panel immediately above it, and write the remainders in the panel immediately below. Carry out the checks as before.

7. Note that, in each hierarchical matrix, the last figure in the criterion column (headed x) will represent the amount of variance contributed by the factor that is being subtracted. Thus, ·3969 represents the variance contributed to the square of the multiple correlation by the first test; ·2653 the variance contributed by the 'reduced' second test (i.e., by the residual component left after the influence of the first test has been removed from the second test); and so on. If at any stage the additional variance appears small, it can thus be tested for significance (by the method described below in Section III) before carrying the calculations further. Should it then seem clear that no significant additions to the variance are likely to be secured by continuing the calculation, the forward solution may be regarded as completed, and the required regressions can be calculated by the 'back solution' from the figures so far obtained: (see Section V).

8. As a rule, however, it will be wiser to continue the calculations until the residual matrix consists of only a single scalar figure. This single figure (here ·3115) will appear in the criterion column, and will represent its specific factor variance. On deducting this figure from 1·0000, the remainder (·6885) will give the square of the multiple correlation coefficient obtained with all four tests.

III. *Determining and Testing the Multiple Correlations.*

9. To calculate the multiple correlation for *all* the tests it will be better to proceed as follows. Add the last figures in the last column from each hierarchy: (·3969 + ·2653 + ·0260 + ·0003 = ·6885). The square root of the sum (·8297) will be the coefficient of multiple correlation required.

10. Since the increments due to the later tests may not be significant, it will be well to add the figures cumulatively as shown in Section D of the table to ascertain how far the addition of each further test improves the correlation so calculated.

11. To test the significance of the multiple correlation, the following ratio should be calculated:

$$F = \frac{R^2 \div m}{(1 - R^2) \div (N - m - 1)},$$

where R denotes the multiple correlation obtained with m constants and N denotes the number of persons tested (here assumed to be 50).

Test of Significance of Multiple Correlation

Source of Variance	Square Sum	D.f.	Mean Square
All four tests	·6885	4	·1721
Residual	·3115	45	·0069
Total	1·0000	49	

$$F = \frac{\cdot 1721}{\cdot 0069} = 24{\cdot}9.$$

With 4 and 45 degrees of freedom this is highly significant.

12. To test the significance of the increment to the multiple correlation obtained by adding one or more constants,[1] the following ratio should be calculated:

$$F = \frac{(R_2^2 - R_1^2) \div (m_2 - m_1)}{(1 - R_2^2) \div (N - m_2 - 1)},$$

where R_1 denotes the multiple correlation obtained with m_1 constants and R_2 the multiple correlation obtained with m_2 constants.

[1] See, for example, Tippett, *Methods of Statistics*, p. 208 and refs. The student may be reminded that it is not strictly accurate to apply a multiple correlation, calculated for a *given* sample, to some *other* sample, even though that other sample is expected to differ from the given sample only by chance: the standard deviation of such estimated coefficients will be larger than the ordinary formula assumes. Hence, in considering the probable value ($\bar{R}$, say) to be obtained in a second sample, it is well to use a correction formula, viz.,

$$\bar{R}^2 = 1 - (1 - R^2)\frac{N - 1}{N - m},$$

where m is the number of constants in the regression equation (Ezekiel, M., *Methods of Correlation Analysis*, 1930, p. 177; Wherry, R. J., *Ann. Math. Stat.*, II, 1931, pp. 440f.). In an empirical study it has been shown that the shrinkage due to the addition of a further variable may at times more than counterbalance the increase gained, when calculated by the uncorrected formula (Larson, S. C., 'The Shrinkage of the Coefficient of Multiple Correlation,' *J. Educ. Psych.*, XXII, 1931, pp. 45–55).

Enter the table of *F*-ratios (given by Snedecor or other available text-book) with the degrees of freedom indicated by the divisors, e.g., in the latter case take $n_1 = m_2 - m_1$ and $n_2 = N - m_2 - 1$.

Test of Significance for an Increment to the Multiple Correlation

Source of Variance	Square Sum	D.f.	Mean Square
First two tests	·6622	2	—
Third test	·0260	1	·0260
Residual	·3118	46	·0068
Total	1·0000	49	

$$F = \frac{\cdot 0260}{\cdot 0068} = 3\cdot 83 .$$

With 1 and 46 degrees of freedom, *F* should be at least 4·05 to reach the 5 per cent. level of significance. Hence the last two tests might well be dropped.

IV. *Re-checking the Order in which the Tests are Taken.*

13. As each test is removed, it will generally be wise to ascertain whether the next test (according to the initial order) is really the best to select. For this purpose, square the last set of residual correlations between the tests and the criterion (i.e., the residuals entered in the column headed *x*), and then divide each square by the residual in the diagonal element in the same row. Select as the next test to be eliminated that particular one which provides the largest ratio.

In Table VIII the columns headed ' squares ' and ' ratios ' give the figures thus obtained for the observed correlation matrix and for each residual matrix. Those entered for the observed correlations (enclosed in brackets) are inserted for completeness only, and have no practical importance : (they would merely mean that we should arrange the tests in the order of their correlations with the criterion). It will be noted that the first ratio in each set reappears as the last figure in the following criterion hierarchy. Like that, it and the other ratios specify the amount of variance contributed towards the square of the multiple correlation by the factor that would be based on each test.

If we turn to Thomson's table (*loc. cit.*, p. 93) where test *a* is taken last, we obtain the following figures (sect. *B*) :

Test	Squares	Residuals	Ratios
c	$\cdot 083^2 = \cdot 0069$	·524	·013
d	$\cdot 057^2 = \cdot 0032$	·760	·004
a	$\cdot 349^2 = \cdot 1218$	·848	·144

These results make it clear that test *a* is contributing a much larger share to the multiple correlation than tests *c* or *d* ; and, as we have seen, the ratio calculated at the outset suggests that its contribution is also larger than that of test *b*.

V. *Calculating the Partial Regressions : Back Solution.*

If the regression coefficients are required as well as the multiple correlation, these can be obtained by a ' back solution ' as follows.

14. Copy down in rows, as shown in Table VIIIB, p. 52, the multipliers from the columns in the left-hand margin of the forward solution (omitting the initial figure, 1·000, in each case), viz., ·3900, ·1900, etc., and then, in the next line but one, ·7264, ·4537, etc., and so on.

15. Now work from the bottom line upwards. The last figure of all, ·0183, taken just as it stands, will be the partial regression for the last test of all, test *d*. Enter this in the column for the regressions on the right.

16. Multiply the figures in the last column but one (those for test *c*) by this figure, and prefix a minus sign. In the last row but one the product will be $- \cdot 0183 \times \cdot 0954 = - \cdot 0017$. Subtract this figure (as indicated by the minus sign) from the last figure in that row : we obtain $\cdot 2242 - \cdot 0017 = \cdot 2225$. This will be the partial regression for the last test but one (test *c*).

17. Multiply the figures in the last column but two (those for test *b*) by this figure, and prefix a minus sign. In the last row but two subtract the products from the final figure. The result will be the partial regression for the next test (test *b*).

18. Continue in this way until all the partial regressions have been calculated.

19. Check the results (i) by multiplying each of the regressions by the corresponding correlations from a row of inter-correlations other than the first, and then summing the products : the sum should yield the corresponding criterion correlation ; (ii) by multiplying each regression by the corresponding criterion-correlation and summing the products : the sum should yield the square of the multiple correlation (see Section C of Table VIII). If the multiple correlation has already been calculated from the forward solution, this provides a check both on the multiple correlation and on the regressions.

Vol. IV
Part I

The British Journal of Psychology,
Statistical Section

March,
1951

THE ESTIMATION OF CORRELATION COEFFICIENTS FROM THE VARIANCE OF SCORE DIFFERENCES

By W. G. EMMETT

Education Department, Moray House,
University of Edinburgh

I. *Introduction.* II. *Necessary Conditions.* III. *Errors Found in Practice.*

I. INTRODUCTION

It often happens that an estimate of a correlation coefficient in a population is required when the scores of a selected sample of restricted range are alone available. If the standard deviations of the two variables, x_1 and x_2, in the population are equal, the required correlation may be rapidly estimated with very small error by finding the variance of the algebraic differences between pairs of scores and applying the formula,

$$\rho_{12} = \frac{2\sigma^2 - s^2_{(1-2)}}{2\sigma^2}, \qquad (1)$$

where ρ_{12} is the estimated correlation in the population, σ^2 the variance of each variable in the population, and $s^2_{(1-2)}$ the variance of the score differences in the sample estimated with $(N - 1)$ degrees of freedom. (Greek symbols are used to denote statistics of the population, and Roman symbols statistics of selected samples.)

The formula follows from the product-moment correlation formula,

$$\rho_{12} = \frac{\sigma_1^2 + \sigma_2^2 - \sigma^2_{(1-2)}}{2\sigma_1\sigma_2}$$

for the case where $\sigma_1 = \sigma_2$, and $\sigma^2_{(1-2)} = s^2_{(1-2)}$. It is therefore assumed that the variance of score differences is the same in the sample as in the population. There is nothing new about the method ; it has been used in Moray House for many years, and was freely applied by G. A. Ferguson (1) in his researches on the reliability of mental tests. The present paper is concerned with an examination of the conditions under which formula (1) is valid and with the errors found in practice for varying departure from these conditions.

II. NECESSARY CONDITIONS

These conditions are now examined. The assumption that the variance of score differences is independent of selection implies that

$$\sigma_1^2 + \sigma_2^2 - 2\rho_{12}\sigma_1\sigma_2 = s_1^2 + s_2^2 - 2r_{12}s_1s_2 . \qquad (2)$$

By Karl Pearson's formula for the effect of selection on a correlation coefficient, we have $r_{12} = \dfrac{\rho_{12} - \sqrt{(1 - p_1^2)\ (1 - p_2^2)}}{p_1 p_2}$, where selection is made through a third

variable, and $p_i = \frac{s_i}{\sigma_i} (i = 1, 2)$. Substituting for p_1 and p_2 and simplifying, we find $r_{12}s_1s_2 = \rho_{12}\sigma_1\sigma_2 - \sqrt{(\sigma_1^2 - s_1^2)\ (\sigma_2^2 - s_2^2)}$. Substituting now for $r_{12}s_1s_2$ in (2), we arrive at

$$\frac{(\sigma_1^2 - s_1^2) + (\sigma_2^2 - s_2^2)}{2} = \sqrt{(\sigma_1^2 - s_1^2)\ (\sigma_2^2 - s_2^2)}\,, \quad (3)$$

that is, the arithmetic mean of two quantities is equal to their geometric mean, a condition that is only satisfied when the two quantities are equal.

Thus

$$\sigma_1^2 - s_1^2 = \sigma_2^2 - s_2^2\,. \quad (4)$$

(I am indebted to Sir Godfrey Thomson for an analysis much the same as the above.)

Equation (4) may be written, $\sigma_1^2 (1 - p_1^2) = \sigma_2^2 (1 - p_2^2)$,

or

$$\sigma_2 = \sigma_1 \sqrt{\frac{(1 - p_1^2)}{(1 - p_2^2)}}\,. \quad (5)$$

It follows that $\sigma_{(1-2)}^2 = s_{(1-2)}^2$ only when x_2 has a standard deviation of $\sigma_1 \sqrt{\frac{(1 - p_1^2)}{(1 - p_2^2)}}$, a condition which can be achieved by multiplying the x_2 scores by this factor. Since p_1 and p_2 vary according to the degree of selection in the sample examined, the s.d. of x_2 relative to that of x_1 will in general differ from one sample to another, unless of course $\rho_{12} = 1$. The re-scaling of x_2 in this manner in the formula

$$\rho_{12} = \frac{\sigma_1^2 + \sigma_2^2 - s_{(1-2)}^2}{2\sigma_1\sigma_2} \quad (6)$$

necessarily leads to the Pearson correction formula, and to the same estimated value of ρ_{12}, as if we had found the sample correlation and then applied the Pearson formula.

III. ERRORS FOUND IN PRACTICE

If in practice the two variables have different standard deviations in the population, no advantage would be gained by changing the scale of one or other variable before forming the distribution of the differences between pairs of sample scores : a direct calculation of the correlation in the sample and the application of Pearson's formula would probably take less time. We are therefore concerned with errors which arise with varying departure from the requirements of equations (4) or (5) in those cases where $\sigma_1 = \sigma_2$, that is, when no correction is applied for the term $\sqrt{(1 - p_1^2)/(1 - p_2^2)}$ in equation (5).

(*a*) We first consider the correlation between two parallel tests. Here the standard deviations of the test scores in the population are approximately the same, indeed are made the same if the scores have been standardized to a common standard deviation as is usual. Further, the correlation between the tests is usually well above 0·9, so that selection will have approximately the same influence on the standard deviation of both variables in the sample. Thus to a first approximation equation (4) is satisfied.

In Table I we give the correlations between two intelligence tests, M.H.T. 43 and M.H.T. 45, which were taken by a complete year group of 11+ children within one education authority at dates 41 days apart. The two tests were standardized independently, each to

a mean I.Q. of 100 with standard deviation of 15. The variance of the differences between each pair of I.Q.'s was found for different samples artificially selected from the year group, and an estimate of the correlation in the population made for each sample from formula (1).

TABLE I. CORRELATIONS BETWEEN TWO INTELLIGENCE TESTS

Sample	N	$s^2_{(1-2)}$	ρ_{12}
(1) Year Group	1,194	18·203	·9595
(2) I.Q.'s 75 to 125	1,059	17·338	·9615
(3) I.Q.'s 101 and above	598	21·771	·9516
(4) I.Q.'s 100 and below	583	14·414	·9680
(5) Primary School A	90	17·748	·9606

Samples (2) to (4) were formed by selecting those children whose mean I.Q.'s lay within the stated limits. Sample (5) represented all boys in the largest school in the district. The population values of the reliability coefficient as estimated from the samples agree very well with the observed value given in the first row of the table. The reliability of the I.Q.'s of the duller half of the year group is greater than that for the brighter half : this accords with other investigations (e.g., Ferguson (1), Chap. XIII, and Weaver (5)).

(*b*) We next consider cases where the variables have the same standard deviations in the population but are less highly correlated. For this purpose we used the scores of a complete year group of 6,348 children of mean age 11 : 2 in three Moray House tests, viz., English M.H.E. 10 (x_1), intelligence M.H.T. 25 (x_2), and arithmetic M.H.A. 10 (x_3), the maximum possible scores in the tests being 150, 100, and 100 respectively. The raw scores were punched on Hollerith cards and various samples drawn with the aid of a mechanical sorter.

First a set of samples was formed by selecting those children whose English scores lay within the different limits shown in column 1 of Table II, and pairing their scores in English and intelligence. Since the variables were measured in units of raw score, in order to test the adequacy of formula (1) it was necessary to change the units for the variables so that each had the same s.d. in the population, a value of $\sigma = 15$ being chosen for this purpose. The sample variances, s_1^2 and s_2^2, were found in terms of raw score and these values were multiplied by $15^2/s_i^2$ ($i = 1, 2$) to effect the necessary change in scale, giving $s_1'^2$ and $s_2'^2$ say. The sample correlation coefficient was then found from a correlation grid, and $s'^2_{(1-2)}$ evaluated from $(s_1'^2 + s_2'^2 - 2r_{12}\, s_1'\, s_2')$, $s'^2_{(1-2)}$ now being referred to a scale of $\sigma = 15$. This indirect procedure was followed as we had already made most of the computations for another purpose ; the result however is exactly the same as if we had started with scores based on a population s.d. of 15 (' quotients ' we call them) without any knowledge of the s.d.'s within the sample or of the sample correlation coefficient, and then found the variance of the difference of the sample scores. The estimated value of ρ_{12} is now given by $\dfrac{450 - s'^2_{(1-2)}}{450}$.

The results are recorded in Table II.

Row 1 in the table gives the statistics for the complete year group ; the correlations estimated from the samples thus have to be compared with ·8883. Samples 2 to 6 have been formed by lopping off varying amounts from one tail of the English score distribution, and samples 7 to 12 by similarly treating both tails. The departure from equality in s_1' and s_2', i.e., from the conditions required by equation (4), can be seen by comparing columns 3 and 4. The population correlations estimated both from the variance of score differences and by Pearson's formula are remarkably close to the directly ascertained value ; in no case is the difference greater than ·01 in spite of the ruthlessness of the selection. Sample 11 is of particular interest on account of the large disparity between s_1' and s_2', and the stringent degree of selection ($p_1 = 5{\cdot}7470/15 = {\cdot}3831$), which results in the low value of ·5802 for the sample correlation coefficient. The population correlation coefficient estimated from the sample is practically the same as that observed.

TABLE II. CORRELATIONS BETWEEN ENGLISH (x_1) AND INTELLIGENCE (x_2)
for varying degrees of selection in English

Sample No.	Range of English Scores (x_1)	No. of Cases	On scale $\sigma = 15$			r_{12}	Estimated value of ρ_{12}	
			s'_1	s'_2	$s'^2_{(1-2)}$		From formula (1)	From Pearson's formula
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	0–150 (popn.)	6,348	15·0000	15·0000	—	—	·8883	—
2	20–150	6,130	13·9242	13·9264	51·482	·8673	·8856	·8856
3	40–150	5,503	12·1529	12·3428	50·779	·8309	·8872	·8870
4	50–150	4,994	11·1294	11·4821	49·429	·8071	·8902	·8898
5	60–150	4,281	9·8968	10·5763	47·547	·7751	·8943	·8934
6	70–150	3,541	8·7316	9·6766	45·997	·7331	·8978	·8966
7	10–129	6,146	14·0041	14·1716	50·705	·8723	·8873	·8868
8	20–119	5,749	12·5908	13·0841	51·319	·8450	·8860	·8845
9	30–109	5,062	10·5639	11·5825	51·136	·7953	·8864	·8836
10	40– 99	4,151	8·2561	10·1041	51·031	·7146	·8866	·8820
11	50– 89	2,994	5·7470	8·6657	50·335	·5802	·8881	·8824

Next a set of samples of English (x_1) and arithmetic (x_3) scores was obtained by selecting children whose intelligence scores (x_2) lay within certain ranges. The variance of the difference in English and arithmetic scores scaled to a population standard deviation of 15 was found by the same method as for the first set of samples ; thence an estimate of the correlation in the population was found from formula (1) as before. The results are given in Table III. Here again the agreement between the estimated values of ρ_{13} and the observed is excellent.

TABLE III. CORRELATIONS BETWEEN ENGLISH (x_1) AND ARITHMETIC (x_3)
for varying degrees of selection in Intelligence (x_2)

Sample No.	Range of Intelligence Scores (x_2)	No. of Cases	On scale $\sigma = 15$			r_{13}	Estimated value of ρ_{13}	
			s'_1	s'_3	$s'^2_{(1-3)}$		From formula (1)	From Pearson's formula
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
12	0–100 (popn.)	6,348	15·0000	15·0000	—	—	·7811	—
13	20–100	5,827	13·2746	13·7225	101·337	·7224	·7748	·7729
14	30–100	5,220	12·2465	12·7585	103·420	·6699	·7702	·7689
15	40–100	4,244	11·1352	11·5218	102·948	·5994	·7712	·7708
16	50–100	2,986	9·7199	10·2415	100·190	·4981	·7774	·7769
17	10– 89	6,133	14·0656	14·5216	100·218	·7552	·7773	·7726
18	20– 79	5,690	12·8942	13·4019	102·342	·7046	·7726	·7706
19	25– 74	5,229	11·9186	12·5912	104·472	·6534	·7678	·7658

(*c*) When the standard deviations of the variables x_1 and x_2 in the population differ appreciably, serious errors in the estimate of ρ_{12} result from the use of formula (6).

As an example the following statistics of the raw scores of Sample 11 of Table II have been taken.

Population : $N = 6{,}348$, $\sigma_1 = 29{\cdot}5728$, $\sigma_2 = 18{\cdot}3576$, $\rho_{12} = {\cdot}8883$.
Sample : $N = 2{,}994$, $s_1 = 11{\cdot}3303$, $s_2 = 10{\cdot}6055$, $s^2_{(1-2)} = 101{\cdot}419$, $r_{12} = {\cdot}58018$.

The estimate of ρ_{12} from the sample data by formula (6) is 1·0224!

If now values of x_2 are multiplied by $\frac{\sigma_1}{\sigma_2}\sqrt{\frac{(1-p_1^2)}{(1-p_2^2)}}$ (= 1·82300) and a new value for $s^2_{(1-2)}$ found by a similar adjustment of x_2 in $(s_1^2 + s_2^2 - 2r_{12}s_1s_2)$, the estimated value of ρ_{12} becomes ·8824. Pearson's selection formula gives exactly the same result, as indeed it must. The estimate of ρ_{12} from this sample given in column 7 of Table II, viz., ·8881, was obtained by multiplying x_2 scores by σ_1/σ_2, i.e., by 1·61093 instead of 1·82300. This correlation happens to lie nearer the observed value than does that obtained from Pearson's formula, but no consequence should be attached to this.

It may be observed that the Pearson correction formula has yielded excellent estimates of the population correlation coefficient, although in many of the samples the distributions have diverged widely from normality. This result agrees with the theoretical views reached by M. S. Bartlett (2) and D. N. Lawley (3, 4).

REFERENCES

1. Ferguson, G. A. (1947). *The Reliability of Mental Tests.* London : University of London Press.
2. Bartlett, M. S. (1938). ' Further aspects of the theory of multiple regression.' *Proc. Camb. Phil. Soc.*, **35**, 180.
3. Lawley, D. N. (1943). ' A note on Karl Pearson's selection formulæ,' *Proc. Roy. Soc. Edin.* A, LXII, 28.
4. Thomson, G. H. (1944). ' The applicability of Karl Pearson's selection formulæ,' *Brit. J. Psych.*, XXXIV, 105.
5. Weaver, H. R. (1950). ' The consistency of I.Q.'s at different levels of ability.' *B. Ed. Thesis, University of Edinburgh.*

BOOK REVIEWS

The Structure of Human Ability. By P. E. Vernon. London: Methuen & Co. 1950. Pp. viii + 160. 12s. 6d. net.

In writing this book, Professor Vernon tells us, he had two principal aims: first, "to present factor analysis simply and show that it has extremely important practical bearings"; secondly, "to show that the large number of publications in this field can be fitted into one consistent, even if incomplete, picture." The first three chapters contain a useful, non-mathematical account of the principles underlying different types of factor analysis. Professor Vernon devotes most attention to three main types, which he illustrates with a clear but simple diagram: (1) Spearman's 'two factor analysis,' (2) Burt's 'group factor analysis,' and (3) Thurstone's 'multiple factor analysis.' So far as cognitive abilities are concerned, he strongly favours the second, i.e., the method which, directly or indirectly, includes both a general factor and a set of group factors. He refuses, however, to identify the general factor, or *g* as he calls it, with the "subjective and indefinable term intelligence," and prefers to define it simply as an objectively established factor running through all the tests in an appropriately selected battery. In this section there are perhaps a few minor points which some readers may find unconvincing or difficult to follow; but that is largely because the discussion is so condensed.

The second and main portion of the work consists of eight chapters describing in some detail the various factors for which there is now fairly adequate evidence. He takes in turn the main fields of statistical research, beginning with early work on educational attainments and ending with recent work on occupational abilities. Thus his second task resembles that of Professor Burt in his early survey of *The Measurement of Mental Capacities* and his more recent review of 'The Results of Factor-Analysis.'[1] Burt's account was based chiefly on work with children; Vernon is able to draw largely on work in the fighting services among adults. The picture that he outlines is very similar to the earlier. He fully agrees that factors are primarily 'principles of classification,' and that the structure so revealed comprises a multiplicity of factors, general and group, broad and narrow, arranged roughly in accordance with what the logicians call a 'hierarchical' scheme, where genera include species, species are divided into sub-species, and so on. The mode of classification thus resulting will be readily understood if the reader turns to the 'pedigree diagrams' printed in earlier issues of this *Journal*.[2]

Until recently, comparatively few factorial investigations have been carried out on large-scale test-results of adults. Now, however, so Professor Vernon assures us, "results obtained from analysis among recruits during the 1939–45 war, have confirmed the hierarchical theory to which this book is committed." Indeed, the list of factors which he enumerates agree in a remarkable fashion with those that originally emerged from factorial work with children.

Professor Vernon does not himself define a 'hierarchy'; he leaves its nature to be gathered from his concrete instances. This perhaps is an omission which the non-

[1] 'The Structure of the Mind: A Review of the Results of Factor Analysis,' *Brit. Journ. Educ. Psychol.* XIX, 1949, pp. 100–11, 176–99.

[2] E.g., Banks, C., 'Factor Analysis of Assessments for Army Recruits,' this *Journal*, II, pp. 84–7. The data on which this analysis was based were obtained from the first 1,000 cases tested at the Primary Training Centres set up by the War Office. Although it was not possible to publish them until after the war, these were, I believe, the first set of army data to be factorized in this way, and the first to show this type of classification among the mental abilities of adults. It is gratifying to learn that the conclusions then drawn have since been so fully confirmed. Among children, of course, the evidence for a hierarchy of subdivided factors goes back to Burt's early work at Oxford. The concept had already been put forward on purely speculative grounds by writers like Spencer, Sully, Binet, and McDougall.

mathematical reader may be tempted to regret, particularly as the broader lines of classification implied by his chapter headings seem to relate to the main branches of research rather than to the main branches of mental ability. For fuller explanations he refers to Burt's recent article on *Alternative Methods ;* but that is a paper which does not seem strictly relevant for the present purpose. Curiously enough, he omits to mention the earlier papers on " The Structure of the Mind " and on " Subdivided Factors," [1] to which the enquiring reader should by preference be sent for a more explicit and detailed account of the underlying theory.

Professor Vernon deals, in a particularly clear, concise, and convincing way, with current criticisms of the main theory and with alternative views. As he points out, " Thurstone and other factorists in America, from 1938 on, opposed this notion of a general factor hierarchy." There are, so he suggests, two main differences between the approaches of British and American factorists : first, " British workers make *g* as large as possible, whereas American writers . . . tend to minimize it," and indeed usually omit it from the primary factors altogether ; secondly, " British factorists recognize larger and more comprehensive group factors, together with sub-factors ' descended ' from them, whereas American primary factors usually possess much the same status and variance." In the Appendix he sets out seven main arguments for rejecting the conception which Thurstone and his followers have chiefly favoured. However, as he points out, during the last few years American writers have been approaching more and more to the British point of view : they now appear, perhaps rather tentatively, to accept a general factor as a factor of second-order, and even at times sub-divide their broader group factors. On these and other controversial points Professor Vernon, as he says, " almost always agrees " with Burt and Thomson, and " usually disagrees " with Thurstone and Spearman.

A special merit of the book is the detailed account of numerous factorial results obtained during the war from tests applied to samples of an exceptionally large size. Many of these war-time analyses, drawn from the Navy, Army, and Air Force, are now published for the first time. It is clear that the conclusions thus obtained have been of great value, both practical and theoretical. They were of definite assistance during the war to psychologists engaged in planning work on personnel selection, and have helped to elucidate the puzzling problem of educational classification for children at eleven plus ; at the same time the increasing measure of agreement which they show seems to be rapidly solving many of the long-standing issues about the theoretical structure of the mind. But, quite apart from its new and original factual contributions, these chapters will also be valued for their critical and comprehensive discussion of factorial researches, particularly from 1935 onwards.

In a book just published [2] Mr. Zwangwill has assured us that " factorial analysis, despite its impressive mathematical procedures, is a brilliant but misguided departure." To this and similar criticisms Professor Vernon's book provides an eloquent and a concrete answer. It is the first in a new series of Manuals of Modern Psychology that are to be issued by its publishers ; and, for its intrinsic interest and critical approach, a better volume could scarcely have been chosen. *E. K. C. B.*

[1] This *Journal*, II, pp. 41–63.

[2] Zwangwill, O. L., *An Introduction to Modern Psychology* (1950). Mr. Zwangwill's own account of factorial work contains nothing more ' modern ' than Spearman's two factor theory and his method of tetrad differences. He finds no difficulty in showing that such factors will take their place " in the limbo of discarded elements of mind."

Statistical Methods in Research. By Palmer O. Johnson. New York : Prentice-Hall, Inc. 1949. Pp. ix + 377. 46s. net.

Dr. Palmer Johnson has attempted to provide a more rigorous exposition of statistical methodology than is to be found in most current works on the subject written for psychologists. In general he relies on a verbal presentation, preferring logical to highly technical mathematical proofs. The long list of references to original papers shows how thoroughly he has searched the statistical literature for theoretical principles and practical techniques which would be of service to the research-worker in the fields of psychology and education. These techniques derive to a great extent from the contributions of writers in this country, like R. A. Fisher, J. Neyman, and Egon Pearson.

The opening chapters start with a brief account of the essential purposes of statistical analysis. Here the writer emphasizes three points that every research student must be taught : first, how to choose the most effective procedure for the problem in hand ; secondly, what are the assumptions on which each procedure rests ; and thirdly, how to test the applicability of those assumptions to the problem or the data to be dealt with. Throughout the student is encouraged to do his own reasoning instead of mechanically applying rule-of-thumb methods taken blindly from some text-book illustration.

The author presents an admirable outline of the Neyman-Pearson approach to the problem of testing statistical hypotheses. The two types of error which may be committed (rejection of a hypothesis which is in fact true and acceptance of a hypothesis which is in fact false) are carefully defined, and the methods employed for limiting these errors fully discussed. Among the examples of statistical tests he includes, not only those to be found in the majority of recent books, but also the Behrens-Fisher method for testing the significance of the difference between two means when the variances are unequal, Bartlett's chi-squared test, the relatively unknown L_1 test for the homogeneity of a set of variances (tables for the L_1 test are printed in an appendix), and Hotelling's test for the significance of the difference between correlations determined from the same sample. In passing it is instructive to observe that Hotelling's test may be generalized to examine the equality of correlation coefficients in any column of a correlation matrix.

A chapter is devoted to methods of point and interval estimation (the latter including both fiducial and confidence intervals), and contains an excellent if somewhat brief discussion of test-reliability. In each example the underlying assumptions of the measure employed are first tested ; and the maximum likelihood estimates of the relevant correlations are substituted for the customary product-moment formula.

A novel feature is the detailed treatment of ranking methods. The introduction of such techniques was due largely to the early work of educational psychologists, and was at the time strongly criticized by statistical referees of the day ; it is only one of the many examples in which working procedures, originally developed by psychologists, are being refined, and eventually finding their way into the repertory of reputable techniques accepted by the pure statistician. Dr. Johnson gives examples of the coefficient of concordance for multiple rankings, the analysis of variance based on ranks, and the various paired comparison statistics. These extensions of the method, most of which are due to the work of Professor Kendall, will be new to many psychological students ; and the account of them might perhaps well have been expanded.

Several chapters are devoted to the analysis of variance and covariance, and to the design of experiments. The various hypotheses to be examined and the appropriate probability models are clearly stated. Examples are appended, showing how the appropriate statistical tests may be derived from maximum likelihood principles. Two or three fairly complicated illustrations are worked out in detail : there is, for example, a full covariance analysis of a $2 \times 3 \times 3 \times 3$ factorial study of data obtained from educational tests. It is noteworthy that, in every analysis, the test for homogeneity of variances is carried out as a matter of routine. These chapters provide one of the clearest expositions of the technique available to the psychologist, far clearer than Fisher's own.

The final chapter deals with problems of multiple regression. A regression analysis is worked out in full. The working, however, employs Fisher's somewhat cumbersome computational techniques ; this means that the complete inverse of the correlation matrix

is calculated on each occasion, but (what perhaps is still more unfortunate) it is nowhere explicitly stated that it is the inverse of a matrix which is being computed.

As a whole the text is extremely lucid. An excellent choice of references is attached at the close of each chapter ; and the questions set for practice are original and stimulating. One or two general criticisms, however, can be made. Although the author may be erring on the safe side, it seems to the reviewer that too many pages are allotted to detailed calculations ; and the space might well have been used to amplify the more general discussions. For example, in discussing 2×2 tables, correction for continuity is not adopted, though a passing reference is made to it. Nor is the exact factorial test (devised by Fisher) referred to. Indeed the whole treatment of chi-square appears rather inadequate. Incidentally, the χ^2-test is not, as stated in a footnote on page 81, an exact test of the significance of difference between two percentages. No clear distinction is drawn between one- and two-tail tests ; in fact, this problem has led the author into actual mistakes. The F-ratio test for the significance of the difference between two variances is used as a one-tailed test on page 82 when in fact a two-tailed test is required. Despite these minor defects, the book can be unhesitatingly recommended to all research workers and to those students who are already acquainted with the rudiments of descriptive statistics. *A. R. J.*

Traditional Number Rhymes and Games. The Scottish Council for Research in Education. London : The University of London Press. 1950. Pp. 144. 5s. net.

This volume is intended to give assistance in the teaching of number to very young children. The panel concerned in its compilation had two aims more particularly in view : first, to supply teachers with informal number material suitable for their youngest pupils, and, secondly, to try to recover for children some of the heritage of Scottish number rhymes and games. The rhymes are classified under various headings, e.g., those in which money, time, or measurement is mentioned, those involving addition, subtraction, or multiplication, those connected with yearly festivals, and so on. Mathematical psychologists, who are also interested in the psychology of mathematics, will find many suggestive facts and curiosities, some of which throw light even on the early evolution of statistical concepts.

We are told that owing to the influence of wireless, the cinema, and television, and the introduction of organized sport, these traditional games are rapidly dying out. The Council has performed a valuable service in issuing a publication which should do much to ensure that such games can be played in the classroom, even if they are disappearing from the playground. *G. K.*

Vol. IV
Part II

The British Journal of Psychology,
Statistical Section

June,
1951

DIFFICULTY FACTORS ARISING FROM THE USE OF TETRACHORIC CORRELATIONS IN FACTOR ANALYSIS

By NEIL GOURLAY

Department of Education, University of Birmingham

I. *Purpose and Plan of Present Study.* II. *Guilford's Investigation :* (*a*) *Description ;* (*b*) *Suggested Alternative Explanation ;* (*c*) *Wider Application of Findings.* III. *Burt and John's Investigation :* (*a*) *Brief Description ;* (*b*) *Suggested Alternative Explanation ;* (*c*) *Significance for Further Investigations.* IV. *Summary and Conclusions.*

I. PURPOSE AND PLAN OF PRESENT STUDY

This paper is an attempt to solve the problem of the ' difficulty factors ' [1] that arise from the use of tetrachoric correlations in factor analysis. As Vernon has pointed out in a recent article (1), there are two well-known researches in which difficulty factors have been obtained from a factor analysis based on tetrachoric correlations : (*a*) an analysis of Seashore's Pitch Discrimination Test carried out by Guilford (2) ; (*b*) an analysis of the Stanford-Binet items for ages X and XII carried out by Burt and John (3). The plan of this paper is to consider these researches in turn, and, in each case, to suggest an alternative explanation for the difficulty factor or factors obtained. The wider implications of the two alternative explanations will be set forth ; and a general hypothesis will be formulated to explain difficulty factors arising from any factor analysis using tetrachoric correlations.

II. GUILFORD'S INVESTIGATION

(*a*) *Description.* Guilford factorized the responses of 300 students to the ten sets of items in the Seashore Test of Pitch Discrimination. His aim was to examine

TABLE I. CENTROID FACTOR-LOADINGS IN THE TEN SUB-TESTS OF THE SEASHORE TEST OF PITCH DISCRIMINATION (GUILFORD)

Stimulus Difference	k_1	k_2	k_3	Communality
30	·76	·41	—·19	·781
23	·79	·35	—·29	·837
17	·86	·24	—·05	·800
12	·80	·29	·21	·774
8	·88	·13	·26	·864
5	·72	—·36	·19	·685
3	·63	—·45	—·15	·622
2	·59	—·29	—·32	·528
1	·12	—·47	·18	·271
0·5	—·14	—·19	—·00	·058

[1] The idea of a ' difficulty factor ' appears first to have been stated and explained in Burt's *Factors of the Mind* (1940), p. 198, and the earlier papers there cited.

the hypothesis " that different abilities are brought into play when items are easy as compared with the situation when items are difficult." Because of the skewness of some of the scores he chose a tetrachoric procedure ; and, using the centroid method, obtained three significant factors. The results are shown in Table I. When the axes were rotated, loadings were obtained as shown in Fig. 1. Guilford had some

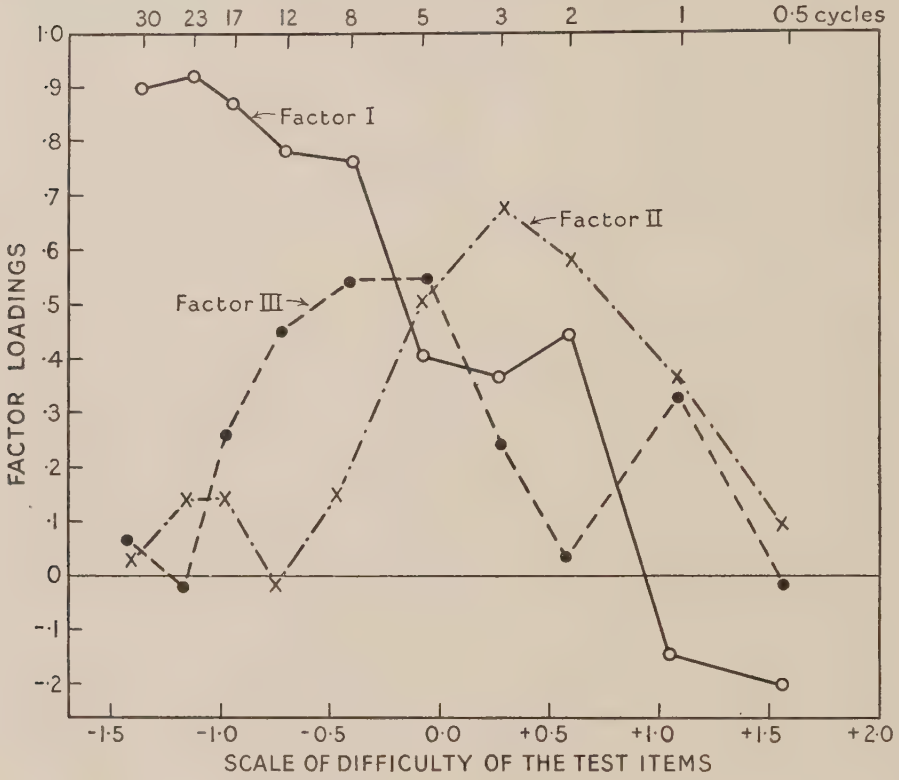


Fig. 1.

trouble in interpreting his factors. He considered the possibility of their being statistical artefacts, but finally accepted the hypothesis that they represented three distinct abilities involved in the comparison and judgment of pitch.

(*b*) *Suggested Alternative Explanation.* The idea underlying the explanation here proposed is to be found in Carroll (6), although not applied specifically to Guilford's investigation. Ferguson (4) had shown that if a *point* correlation is used to factorize items of the same psychological content but of different difficulty, spurious difficulty factors are obtained. Wherry and Gaylord (5) further claimed that, by using the *tetrachoric* correlation, one can avoid such factors. Carroll qualified the conclusions of Wherry and Gaylord : he showed that, when the items analysed are of the multiple choice type (a common type in mental tests), the chance element may operate to produce a new type of spurious factor, even when the tetrachoric procedure is employed. I was unaware of Carroll's article when I first produced my solution of Guilford's problem : Carroll's work was also overlooked by Vernon (1).

Consider a hypothetical set of eight items of the same psychological content, and let their level of difficulty (defined by the fractions of the population passing the items) be ·8, ·7, . . . , and ·1. Then, assuming for simplicity that there is no attenuation, the tetrachoric correlation for any pair of items should be unity, since all items have the same content.[1] If a fourfold table is drawn up (see Table II A), showing the proportion of passes and failures, the proportion who fail the easier item but pass the harder will be zero, i.e., $d = 0$.

TABLE II. FOURFOLD DISTRIBUTION FOR TWO ITEMS OF SAME CONTENT (ATTENUATION ZERO)

(A) Without Guessing

Easier Item	Harder Item: Fail	Harder Item: Pass
Pass ..	b	a
Fail	c	$d = 0$

(B) With Guessing (1 in 2 chance)

Easier Item	Harder Item: Fail	Harder Item: Pass
Pass ..	$\frac{b}{2} + \frac{c}{4}$	$a + \frac{b}{2} + \frac{c}{4}$
Fail ..	$\frac{c}{4}$	$\frac{c}{4}$

Now let the items be arranged on the multiple-choice plan ; and consider the special case of items of the true-false type—the type that occurs in Seashore's test. Then testees who do not know the right answer for an item have a 50–50 chance of guessing it. What happens to the fourfold table ? By simple probability theory, assuming the population infinite or very large, we get the distribution given in Table II B. A tetrachoric correlation can then be found for this distribution. Obviously it will be less than unity : the true correlation has been attenuated by the chance factor.

Here the reader may legitimately object to estimating tetrachoric correlations for multiple-choice items without making allowance for chance. But at this stage it could be argued that, for the purpose of factor analysis, as we do not normally make an allowance for ordinary attenuation, there should be no correction for attenuation due to chance. Tetrachoric correlations were therefore found for the eight true-false items (defined above) : computing diagrams were used (7). The correlation matrix so obtained is shown in Table III. It is obviously not hierarchical. Since the correlations are derived from an infinite population, sampling error is not involved. Further, the table possesses a striking feature, which Guilford noted in his own table for the Seashore items : items of similar difficulty correlate more highly with one another than items of dissimilar difficulty.

TABLE III. CORRELATION MATRIX FOR HYPOTHETICAL DATA

Item	a	b	c	d	e	f	g	h
a (·8)		·597	·480	·396	·299	·221	·145	·071
b (·7)	·597		·547	·441	·340	·249	·160	·079
c (·6)	·480	·547		·486	·378	·275	·181	·090
d (·5)	·396	·441	·486		·414	·300	·200	·097
e (·4)	·299	·340	·378	·414		·333	·222	·101
f (·3)	·221	·249	·275	·300	·333		·241	·116
g (·2)	·145	·160	·181	·200	·222	·241		·129
h (·1)	·071	·079	·090	·097	·101	·116	·129	

Difficulty levels (corrected for chance) are shown in brackets.

The results of the centroid analysis for the hypothetical data are shown in Table IV. Theoretically it is possible that eight factors may be required ; but, since the coefficients supplied by the computing diagrams are accurate to two places at most, no more than three factors were extracted. These will account for practically all the variance. The resemblance between this analysis and Guilford's (see Table I) is most marked : there is the same variation of factor-loadings with item difficulty.

[1] The assumption of normality which underlies the use of tetrachoric correlation need not be introduced here. It will be discussed later in the paper.

TABLE IV. CENTROID FACTOR-LOADINGS OF HYPOTHETICAL DATA

Item	I	II	III	h^2
a (·8)	·654	·337	—·195	·580
b (·7)	·711	·335	—·111	·630
c (·6)	·694	·195	·108	·531
d (·5)	·651	·021	·197	·463
e (·4)	·576	—·147	·188	·388
f (·3)	·476	—·278	·034	·305
g (·2)	·347	—·286	—·088	·210
h (·1)	·182	—·177	—·133	·083
Variance per cent. ..	31·89	5·95	2·03	39·87

As a further confirmation of the resemblance between the two analyses, three rotations were made, and new loadings were obtained as shown in Fig. 2. The resemblance of the variation in loadings to that for Guilford's analysis (Fig. 1) is most pronounced, especially if we miss out the two most difficult of the Seashore items. In view of this close similarity, I would suggest that the results obtained by Guilford's analysis are not the effect of the operation of three abilities in pitch discrimination, but merely statistical artefacts.

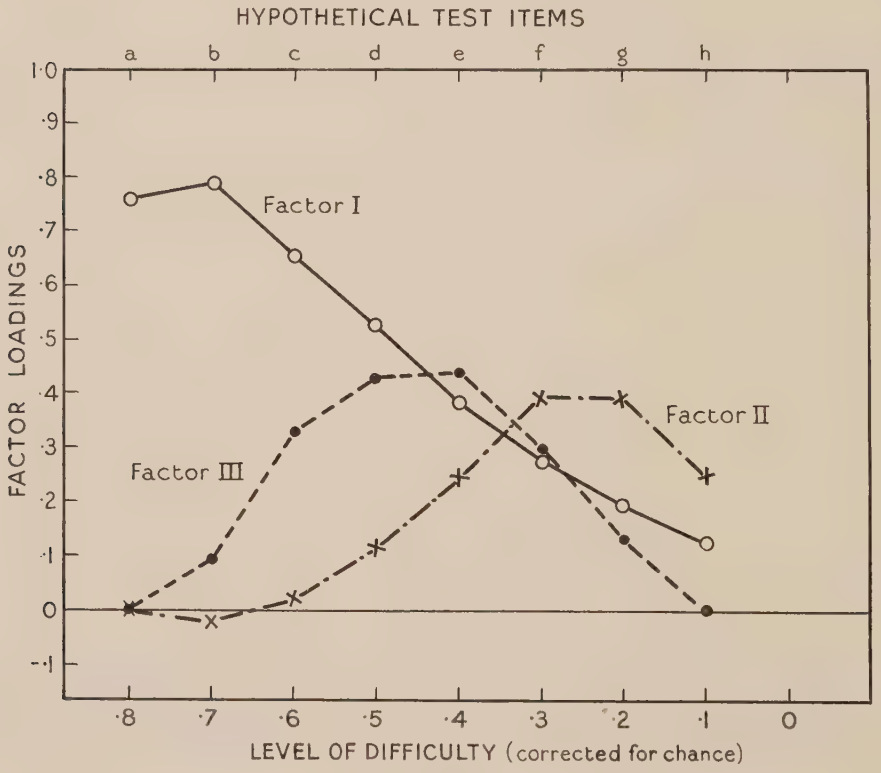


FIG. 2.

(*c*) *Wider Application of Findings.* The wider application of the above argument is apparent. In the factorial analysis of multiple-choice items, where tetrachoric

correlations are employed, the coefficient should be corrected for attenuation due to chance. Provided certain conditions are satisfied (for example, that every item is attempted by all testees), this is easily accomplished by the use of simple probability theory. Carroll outlines and illustrates the method. The same principles underlie Johnson's K.G. Index (8), which is an index of item validity providing a correction for chance success. It will be understood that the smaller the probability of guessing the right answer by chance, the less need there will be for any correction procedure.

III. BURT AND JOHN'S INVESTIGATION

(*a*) *Brief Description.* Burt and John's investigation (3) consisted of a factorial analysis of the X- and XII-year items of the Stanford-Binet Test. The data for the analysis were obtained by choosing a group " as homogeneous as possible in respect of general intelligence, i.e., one having a comparatively narrow range in respect of mental age and mental ratio."

Tetrachoric correlations were calculated for each pair of the twelve items involved. Both a ' bipolar ' factor analysis and a group-factor analysis were carried out. Here we shall consider only the former, the results of which are reproduced in Table V. The interesting feature of this analysis is the first bipolar factor which with only two exceptions separates all the X-year items from the XII-year items: it yields negative saturations for all the tests in age X, except Memory Drawing, and positive saturations for all the tests in age XII, except Repeating Five Numbers Backwards. (The reader should perhaps be reminded that Terman's revision contained no tests for age XI.)

TABLE V. ANALYSIS OF STANFORD-BINET ITEMS (BURT AND JOHN)

Factors	I	II	III	IV
Memory Drawing (X, 3)	·310	·465	—·645	·112
Ball and Field (XII, 3)	·158	·068	—·299	—·400
Fables (XII, 5)	·364	·292	·303	—·308
Mixed Sentences (XII, 4)	·693	·326	·111	—·146
Abstract Words (XII, 2)	·549	·686	·229	·405
Vocabulary (XII, 1)	·724	·131	·302	·339
Vocabulary (X, 1)	·745	—·184	·121	·370
Comprehension (X, 5)	·271	—·335	·133	—·271
Reading (X, 4)	·767	—·412	·435	—·154
Absurdities (X, 2)	·886	—·297	·071	—·237
Repeating Six Numbers (X, 6)	·098	—·143	—·336	·181
Repeating Five Numbers Backwards (XII, 6)	·163	—·597	—·425	·109
Variance per cent.	30·0	14·0	10·6	7·4

Burt's [1] explanation of this bipolar factor is that, since factors are in the first instance abstract principles of classification rather than concrete intellectual abilities, we have here " a mode of classification which distinguishes tests and children more specifically according to their mental age level—we say ' more specifically,' . . . because the crude mental age depends in part on innate intelligence which is measured by the mental ratio ; in part, however, it depends on maturity, that is, roughly speaking, on chronological age, and on the increased development and particularly on the increased acquisition of knowledge and skill that chronological age brings with it."

[1] For brevity the name of only the senior author member will be quoted.

The factor is thus said to be a ' difficulty factor ' of the type described in *Factors of the Mind.* This interpretation, it is argued, explains " why the test of repeating numbers backwards is classed by the bipolar factor, not with the twelve-year-old set (to which it is assigned in virtue of its relative difficulty) but with the ten-year-old set, while the test of copying drawings is classed with the twelve-year-old set and not with the ten-year old ; the latter obviously depends largely on acquired skill, whereas the former does not." It may be noted here that Burt thought his interpretation might also serve to explain a peculiar factor obtained by Wright (9) in her factorial analysis of the Stanford-Binet test : she tested 456 children of chronological ages ranging from ten years to ten years eleven months. The resemblance is not close. Wright's factor does not separate the test items according to age level, but merely gives significant positive loadings to a few items lying at or below the X-year level ; also, as Burt points out, it emerges only after a doubtful rotational procedure, whereas Burt's factor emerges in the initial analysis. Nevertheless, there may be some connexion between the two factors ; and whatever explains Burt's factor may also explain Wright's.

(*b*) *Suggested Alternative Explanation.* Let us first ask how it is that chance has produced ' difficulty factors ' in the Guilford analysis. To answer this question, we must remind ourselves that, in using tetrachoric correlations for a factorial analysis, two things are assumed :

(i) That the distribution of the population relative to the ability underlying performance in any two items is a normal bivariate distribution, and that when we pass and fail members of the population in these two items we are carrying out a double dichotomy of that distribution ;

(ii) That, when we calculate a tetrachoric correlation from the fourfold table resulting from these two dichotomies, the value is independent of the points where the dichotomies are made : i.e., different dichotomies should give the same value, that of the normal product-moment correlation.

In Guilford's experiment, it might well be that the population considered was normally distributed for pitch discrimination. But the same could not be said for the distribution underlying a Seashore test item when no correction has been made for chance : by allowing chance to operate, we are effectively superimposing an error variable on one part of a normal distribution without doing the same to the other part, and the result will be non-normality. Thus, by not correcting for chance, Guilford has introduced non-normality into his data. The conditions specified above were not satisfied ; and difficulty factors resulted.

In Burt's case, we cannot invoke chance to explain the factor. But it might well be that he transgressed the above conditions. Let us see how this is possible.

The assumption of normality is always to be adopted with caution. In more exact sciences such as biology, deviations from normality are the rule rather than the exception (10, p. 132). It is possible that the situation is no different in psychometrics ; but, because of the lack of an accepted unit of measurement in the absolute sense, it is generally agreed that it is impossible to detect any ' real ' deviation from normality.[1] In fact, where the sampling is well done, most psychometrists are prepared to assume normality or near-normality and to define their unit accordingly.

Normality is seldom the case when a sample consists of groups taken from different populations, e.g., children of different age, sex, etc. Even when the separate groups are normally distributed, the combination may deviate markedly from normality (11, pp. 219–28). Now Burt's sample included 483 boys and girls of mental ages between 10 and $11\frac{1}{2}$, of chronological ages from 10 to $14\frac{1}{2}$, and of mental ratios from 80 to 110. Hence non-normality seems quite possible in such a composite group.

Let us consider a non-normal bivariate distribution. Suppose we have for two items a bivariate distribution differing from the normal in having a greater frequency density near the centre : the distribution would be leptokurtic. In Fig. 3 an ellipse is drawn to represent a normal bivariate distribution, and a shaded area is shown for the addition just mentioned. Then as Burt's sample includes mental ages from 10 to $11\frac{1}{2}$ and the items are for years X and XII, the dichotomies must be respectively below and above the 50–50 dichotomy for the sample. Let similar dichotomies be made for the two items in Fig. 3 ; and suppose that the extra density is entirely contained in the rectangle

[1] I do not share this opinion : see Sect. III (*c*) (iii) of this paper.

so formed. What will happen if we now calculate tetrachoric correlations with different pairs of dichotomies? Obviously, dichotomies at the same level will give high correlations, and dichotomies at different levels low correlations. This is the characteristic of Burt's correlation matrix, which leads to the separation of 10 of his 12 items according to age level. A distribution such as the foregoing could therefore provide a possible explanation for this separation as it appears in Burt's first bipolar factor.

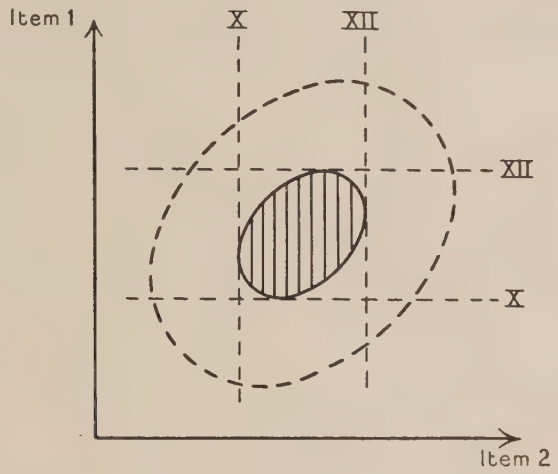


FIG. 3.

How could such a distribution come about? Before answering, let us dispose of two small points. First, Burt had difficulty with the occasional negative coefficients which appeared in his correlation matrix. For his group-factor analysis he treats them as due to the "eccentricities of sampling." But earlier he explains them as possibly due to the differences in performance of older and younger children. Thus an item where the younger do better than the older might correlate negatively with an item where the older have the advantage. But surely another factor is the difference between children, whether young or old, in verbal and non-verbal items. Actually most of the negative correlations are between verbal and non-verbal items; and with a group of a narrow range of intelligence such as Burt took, a verbal item might obviously correlate negatively with a non-verbal item. Secondly, a point which indicates a possible *tour de force* in his argument, Burt explains his first bipolar as a kind of maturity factor, and regards his second as verbal, numerical, and spatial, whereas it could be argued that the second bipolar separates the verbal from the non-verbal items, and the first divides the verbal according to age level and the non-verbal according to a spatial-numerical category. However, as Burt himself has pointed out, in most factorial analyses of groups of tests similar to those considered here (Wright's, for example), the first bipolar usually is the verbal *versus* non-verbal factor; and the fact that it is not in this case deserves careful consideration.

Now let us turn to Burt's explanation of his maturity factor. After stating that his first bipolar simply classifies the tests according to age level, he goes on to explain why the two exceptions, Memory Drawing and Repeating Numbers Backwards, find themselves in the opposite groups: the latter, he says, is put with the X-year items because it depends less on experience and is therefore easier for the children who are chronologically younger, but brighter mentally as a result of their innate intelligence, and the reverse is the case for Memory Drawing which depends more on skill acquired at a later age. If this is so, can we also say the same about the other ten items—that the five which are in the X-year group are easier for younger children and the other five easier for the older ones? I am of the opinion that Burt's argument demands this. Consider the two Vocabulary items (X, 1 and XII, 1) which are classified by the maturity factor according to their stated age level. Is it possible for the older children to do better at the XII-year Vocabulary item, and the younger better at the X-year Vocabulary item? To answer this we must remember that the *same* word list is used at both ages; and, if Terman's instructions are followed, the child who

passes the XII-year item must pass the X-year item.[1] There seems only one solution ; the older pupils must have a greater spread in vocabulary ability than the younger (see Fig. 4). This, of course, is quite possible. It is generally accepted by psychologists that children show progressive differentiation as they approach puberty. Burt himself speaks of this in a footnote to his article (3, p. 119). If so, what would follow? Obviously there will be a greater density about the centre of the total distribution than the normal curve would warrant ; in fact, if the same holds for all the other verbal items, we shall get bivariate distributions of a type which we know must give the kind of factor Burt obtains. This need not be the only influence producing a non-normal distribution. Sex difference, or some other unknown feature of Burt's sampling, may play a part.

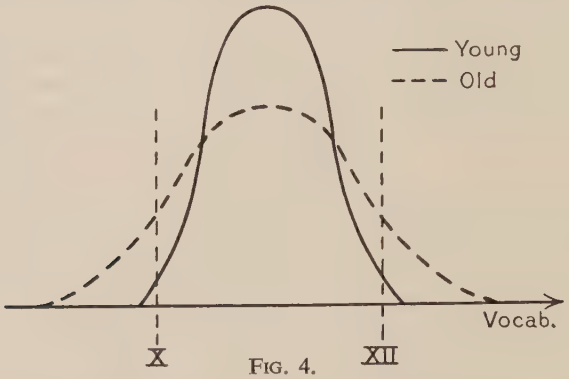


FIG. 4.

There still remains the problem of showing why the two numerical and the two spatial test items were allocated by Burt's maturity factor in the manner found. It will suffice to say that it is not difficult to derive other non-normal bivariate distributions for these items, which would explain the results obtained.

(*c*) *Significance for Further Investigations.* (i) If the assumption of normality has to be made for purposes of tetrachoric correlation, the investigator must choose a population likely to have such a distribution, and draw *random* samples from this population. He should avoid mixing ages, sexes, schools, etc. ; or, if he cannot, he should employ techniques such as those outlined by Lindquist (11), provided always that the necessary assumptions (homogeneity, etc.) are fulfilled.

(ii) I maintain that the appearance of difficulty factors in any factorial analysis based on tetrachoric correlations indicates non-normality somewhere in the data. This is perhaps rather a sweeping conclusion to draw from only two investigations (Burt's and Guilford's). But this hypothesis appears plainly to follow from the second of two assumptions on which the tetrachoric correlation is based. It is interesting to note that in Vernon's study (1) there is some evidence for a difficulty factor resulting from the fact that some items were not attempted by all testees. Once again we seem to be dealing with the case of non-normal distributions. Vernon's analysis therefore provides further confirmation.

[1] It follows that the fourfold table for the two Vocabulary items should have a frequency of zero in one cell, and this in turn would disturb the normality of the bivariate distribution. Unless some allowance has been made for this, the value of the tetrachoric correlation for the two Vocabulary items should therefore be unity and not ·646 as given by Burt and John. Wright shows a similar discrepancy : she had three Vocabulary items in her analysis and gave a value of ·85 for each of the three tetrachoric coefficients involved. Different methods were used in the two cases for the calculation of the coefficients : Burt and John used the full Pearsonian method, whereas Wright used the computing diagrams of Chesire, Saffir, and Thurstone.

Quinn McNemar (*Revision of the Stanford-Binet Scale*, 1942) criticizes Wright for using recurring items although scored differently : " the performance on only one test situation, though scored differently, cannot be regarded as yielding two experimentally independent variables, and of course the correlation between the two will be spuriously high." Shortly afterwards he states that fourfold tables with a frequency of zero in one cell lead to tetrachoric correlations which are greater than unity. I do not agree with this : the value of the tetrachoric correlation in such cases should be unity *exactly*.

(iii) A tetrachoric correlation technique might even be used to detect deviations from normality in psychological data, thus solving a problem generally considered insoluble. However, any demonstration of this must be left to a future paper.

IV. SUMMARY AND CONCLUSIONS

1. Two researches are considered in which difficulty factors appear to result from the use of tetrachoric correlations in a factor analysis. In each case reasons have been given for doubting the hypothesis, and an alternative explanation suggested.

2. In Guilford's investigation, where the items are of the true-false type, the difficulty factors may be explained as spurious factors resulting from the operation of chance. It is concluded that in any analysis of multiple-choice items, difficulty factors will be present unless a correction is made for chance.

3. In Burt and John's analysis, it is suggested, the difficulty factor arose from the failure of the data to satisfy the condition of normality which the tetrachoric procedure assumes : the factor is thus an artefact. The non-normality appears due to the greater differentiation among the older children and possibly to some other feature of the sample.

4. It is suggested that the appearance of difficulty factors in any factorial analysis using tetrachoric correlations is an indication of non-normality somewhere in the data.

REFERENCES

1. Vernon, P. E. (1950). ' An application of factorial analysis to the study of test items.' *Brit. J. Psych., Stat. Sec.*, III, 1–15.
2. Guilford, J. P. (1941). ' The difficulty of a test and its factor composition.' *Psychometrika*, VI, 67–77.
3. Burt, C., and John, E. (1942). ' A factorial analysis of Terman-Binet tests.' *Brit. J. Educ. Psychol.*, XII, 117–27, 156–61.
4. Ferguson, G. A. (1941). ' The factorial interpretation of test difficulty.' *Psychometrika*, VI, 323–9.
5. Wherry, R. J., and Gaylord, R. H. (1944). ' Factor patterns of test items and tests as a function of the correlation coefficient : content, difficulty, and constant error factors.' *Psychometrika*, IX, 237–44.
6. Carroll, J. B. (1945). ' The effect of difficulty and chance success on correlations between items or between tests.' *Psychometrika*, X, 1–19.
7. Chesire, L., Saffir, M., and Thurstone, L. L. (1933). *Computing Diagrams for the Tetrachoric Correlation Coefficient.* Chicago : University Bookstore.
8. Johnson, A. P. (1947). ' An index of item validity providing a correction for chance success.' *Psychometrika*, XII, 51–8.
9. Wright, R. E. (1939). ' A factor analysis of the original Stanford-Binet scale.' *Psychometrika*, IV, 209–20.
10. Kendall, M. G. (1945). *The Advanced Theory of Statistics*, I. London : Griffin and Co., Ltd.
11. Lindquist, E. F. (1940). *Statistical Analysis in Educational Research.* New York : Houghton Mifflin Co.

A REPLY TO MR. GOURLAY'S CRITICISMS

Mr. Gourlay's criticisms of our hypothesis of ' difficulty factors ' raise a number of suggestive questions, each of which can only be answered very briefly here.

1. *Guilford's Investigation.* Guilford's factors do not seem to be ' difficulty factors ' in our sense of the word. They represent ' abilities ' ; and his conclusion is that the ' abilities ' (i.e., mental content factors) which are brought into play with the more difficult items might have been different from those which are involved when the items were easy, even though both sets of items have apparently the same physical content. However, we recognize that his figures could also be interpreted as involving a ' difficulty factor ' in our sense.

We readily endorse Mr. Gourlay's warnings about the use of tetrachoric correlations in such cases. But we cannot agree that the emergence of ' difficulty factors ' is *necessarily* suspect. The common prejudice against such factors seems to arise from the old assumption (expressed, for example,

in the definitions given by both Spearman and Thurstone) that a factor must always denote some ' psychological ability,' i.e., a ' content factor.' Now, unless special precautions have been taken to select items of equal difficulty, the raw figures in all such investigations (i.e., the observed measurements or frequencies) must inevitably include a ' factor ' representing the differing degrees of difficulty in the several tests. But, in preparing those figures for the purpose of correlating traits, the computer usually endeavours to eliminate this factor at the very start, on the ground that it is irrelevant. With measurement data it is eliminated by reducing the initial measurements to deviations about the mean for each test ; with qualitative data the tetrachoric and the point correlation make a similar conversion implicitly. At times, however, particularly with the cruder fourfold tabulation, the elimination may be far from complete. We suspect that this may have been the case in the investigation by Guilford to which Mr. Gourlay refers (Sect. II (*a*), p. 65 above).

In our view, however, the real question is not whether the factor of difficulty is present or absent, but whether it is desirable to treat it as a set of weighted means (as we do in factor analysis) or as a set of unweighted means (as we do when eliminating it by reducing data to deviate form). As Burt has stated elsewhere, " the disadvantage of the latter procedure is that we are discarding information that might be relevant to our main problem ; the advantage of the former is that it enables us to maximize that information."

2. *Gourlay's Interpretation.* We do not think the argument used by Mr. Gourlay in his ' alternative explanation ' of Guilford's results (Sect. II (*b*)) is at all convincing. Mr. Gourlay begins by assuming that his own ' hypothetical items ' differ in difficulty ; he then presents a table of centroid factors (Table IV), the first of which, as it turns out, ranks those items (with one exception) in the precise order of difficulty which he assumes to exist ; and from this he draws the rather surprising conclusion that this factor is *not* a factor of difficulty, but merely a ' statistical artefact.' This surely is the very reverse of the inference his figures would naturally suggest. No doubt the *supplementary* factors in his table (factors II and III) might be regarded as statistical artefacts. But they are due, not to the use of the tetrachoric method, but to the use of the centroid method when, by hypothesis, there were no differences of content in the items selected.

3. *The Binet Investigation.* Mr. Gourlay suggests that " in Burt and John's analysis the difficulty factor " (or what seems to be such) " arose from failure to satisfy the condition of normality." Actually, however, the distributions, when plotted, were found to be as nearly normal as such distributions generally are ; and we made special adjustments to avoid zero-frequencies. Hence Mr. Gourlay's suggestion would certainly not account for the *whole* of the factor he criticizes. We were by no means unaware of the influences to which he refers. He seems to have overlooked the two footnotes where we twice insisted that a difficulty factor, when obtained in this way, " *may* be an artefact " (pp. 122 and 156). With this possibility in mind we looked round for other evidence which might help to decide the point. In the earlier memorandum, in which Burt first suggested applying the tetrachoric procedure to psychological data, he emphasized that it would be safer, wherever possible, " first to cast the tests in an ' internally graded ' form " ; and added that this could be done, and had been done, for most of the problems in the Binet scale. In the later article which Mr. Gourlay quotes, we expressly noted that further evidence had been obtained along these lines ; and, in our view, the strongest reason for rejecting the idea that the factor might be an artificial effect of the tetrachoric procedure was the discovery (mentioned on p. 122) that, when we marked responses on a graded scale and used product-moment coefficients, much the same kind of factor emerged. We therefore feel that Mr. Gourlay's arguments are far from convincing.

4. *The Principle of Maximum Information.* A colleague who has read both Mr. Gourlay's paper and the foregoing comments thinks we should accept Mr. Gourlay's conclusions, while rejecting his arguments. " Mr. Gourlay " (he says) " seems to have raised two questions of great importance. (i) Is it valid to claim that the difficulty of test items can be represented by a factor ? (ii) In whatever way we represent it, do not the customary methods ignore the effects of guessing ? So far as the Binet tests are concerned, surely the relative difficulty of the several items is properly indicated, not by a ' factor,' but by the way they are scaled. In his *Notes on the Theory of Test Construction* Burt attempts to show how the 50 per cent. criterion adopted for scaling purposes is deducible from Fisher's principle of maximizing ' information.' But, if what Mr. Gourlay says about Guilford's data is also true of Burt's data, then surely Fisher's method is no longer applicable. Mr. Gourlay himself seems to concede too much when he says that ' in Burt's case we cannot invoke chance to explain the factor.' Many Binet tests, e.g., ' Distinguishing Right and Left ' or ' Morning and Afternoon ' and ' Repeating 5 or 6 Numbers,' are surely questions to which a correct reply may quite easily be given by blind guessing. If so, two further conclusions may be drawn, namely, (*a*) that, in order to allow for guessing, the tests should be assigned to mental ages, not according to Burt's 50 per cent. criterion, but according to a somewhat higher percentage, such as Terman, McNemar, and Binet himself assumed ; and (*b*) that a factor analysis (which in my view as in Gourlay's should be concerned solely with content) ought to have been based on a still more homogeneous and a decidedly brighter group. Moreover, if guessing took place, we can explain Burt's difficulty factor as an artefact without assuming that there must necessarily have been departures from normality. Note too that Burt apparently admits Mr. Gourlay's main point, namely, that, in such cases, purely artificial factors *may* arise. He thinks the first of his own bipolar factors is genuine. But what

about his *other* factors? And if he admits that, in Gourlay's correlation table, the bipolar factors are artificial, how does he know that the bipolar factors extracted from his *own* table are not equally artificial? Finally, if (as he contends) a centroid or summational analysis is not appropriate in the former case, what method would he propose?"

Once again, owing to limitations of space, we must confine ourselves to commenting very briefly on the main points raised. (i) To scale tests according to their relative difficulty is precisely the same as attempting to assess their saturations with the difficulty factor. This is shown by Burt in his discussion of the difficulty of the Binet tests (*Mental and Scholastic Tests*, p. 452). The advantage of a formal factorial procedure lies simply in the fact that it enables us to obtain the best weights. Ideally these weights should be obtained by the method of maximum likelihood, since that method (as Fisher has demonstrated) "will always extract the maximum amount of available information." (ii) If the Binet test items which involve multiple choice had been given in *group* form, then chance guessing would undoubtedly have operated, and might even have accounted for part of the alleged 'difficulty factor.' But, even so, we cannot agree that 'Fisher's principle' would no longer be applicable. Fisher's procedure is to choose that proportion for which "the quantity of information is greatest," and this, he explains, "may be obtained by a simple application of the differential calculus" (*Design of Experiments*, pp. 226f.). Now, as a matter of fact, such a procedure (as the *Notes on Test Construction*, cited by our critic, proceed to point out) can quite readily be applied to tests involving multiple choice; and, where guessing occurs, a very simple algebraic deduction will indicate the requisite increase in the percentage, which depends simply on the number and likelihood of the choices. But with the Binet tests this problem hardly arises. They are applied individually; and blind guessing is then ruled out, or at any rate immediately detected and allowed for. Moreover, at the higher ages (such as those with which our investigation was concerned) few of the problems really involve 'multiple choice' in the technical sense. (iii) We readily agree that, in theory, it would have been preferable to work with a group that was still more homogeneous; but we were confined to the data we happened to possess. On the other hand, we cannot agree that the groups selected should have been 'decidedly brighter.'

5. *Alternative Methods of Factor Analysis.* Our critic's final question requires fuller discussion. What method of factor analysis should be used with cases like Table III, if the centroid or summational procedures are inappropriate? Here as elsewhere, the method of factor analysis to be selected must depend on the hypothesis that it is desired to test. Mr. Gourlay's hypothesis may be stated as follows: (*a*) the items whose observed correlations are shown in Table III differ solely in degree of difficulty, and not in content; (*b*) every child who possesses the skill needed to answer a harder question necessarily possesses the skill needed to answer all the easier items; (*c*) the only influence operating to obscure the effects of these tendencies is that which arises from random guessing. Now the first two assumptions imply that the composite abilities required by his eight imaginary items are built up of positive group factors related in the following way: the first and broadest factor of all must represent the skill needed to answer the easiest item (this factor will be common to every one of the items, since by hypothesis the capacity to answer a harder item implies capacity to answer the easiest); the next broadest factor will represent the *additional* skill needed to answer the easiest item but one (this factor will not appear in the easiest of all, but will be common to all the rest); and similarly with the remaining group factors. In the end, it will be seen, the n test items will require n factors, each factor covering one item less than its predecessor. Such a set of factors will form a *triangular* factor-matrix. Hence the triangular method of group factor analysis will provide the most appropriate procedure. If the hypothesis is wrong, then, after the n hierarchies have been subtracted, significant residuals will remain. Note however that, as a result of (*b*), the correlated variables will not represent wholly independent observations.

Now, even if Mr. Gourlay had told us nothing about the way he had constructed his Table III, an experienced factorist would, we imagine, think first of applying this procedure, as soon as he had glanced at the table. Its peculiarity is the presence of what Burt has called 'a diagonal ridge'; and this is always a cue for attempting the triangular method of analysis. With such an analysis the first set of factor-saturations obtained from Table III (·79, ·75, ·60, ·50, ·38, ·28, ·18, ·09) proves to be almost identical with his hypothetical figures for difficulty, shown in brackets against each item.

Nevertheless, with the correlations given in his table, small residuals will inevitably remain; and (assuming the sample to have been sufficiently large) these might easily prove statistically significant. There is consequently something wrong with the hypothesis. In explaining it Mr. Gourlay says: "the assumption of normality which underlies the use of the tetrachoric correlation need *not* be introduced here"; and, in constructing both Tables II A and II B, he has actually committed himself (as will be obvious on reflection) to a point-distribution. And yet the correlations that he gives in Table III are calculated on the assumption that the distributions *are* normal, and not point-distributions. Thus, his table of correlations, which we are asked to treat as observed values, really contradicts the initial hypothesis which we are required to test.

If therefore we desire to adhere strictly to the original assumptions adopted in constructing Table II B, we must substitute point correlations for the tetrachoric. In that case, however, the correlations required for the hypothetical model can be deduced directly from the assumed percentages of successes and failures in the following way. Let us write p_a and p_e for the proportionate

numbers passing the more difficult items (d) and the easier items (e) (so that Mr. Gourlay's $a = p_d$ and his $c = 1 - p_e = q_e$ say). If there is no guessing, the point correlation will then be

$$\frac{p_d q_e}{\sqrt{\{q_d q_e(1 - q_d)(1 - p_e)\}}} = \sqrt{\frac{p_d q_e}{p_e q_d}}. \qquad \text{(i)}$$

If there is guessing, it will be

$$\frac{\frac{1}{2} p_d q_e}{\sqrt{\{q_d q_e(1 - \frac{1}{2} q_d)(1 - \frac{1}{2} q_e)\}}} = \frac{p_d q_e}{\sqrt{\{(1 - p_d^2)(1 - p_e^2)\}}}. \qquad \text{(ii)}$$

Evidently both formulæ can be expressed as the product of two saturations, each depending directly or indirectly on the proportion passing one of the tests. These saturations will consequently yield a product table that *is* hierarchical (i.e., a matrix of rank one), though it will not be symmetric. The products in one triangle of the hierarchy will represent the 'observed' correlations (the meaning of those in the other triangle does not concern us here). If now we make our correlation table symmetric by inserting the former set of values in the second triangle as well as the first, and then subject the entire matrix to a factorial analysis by the triangular procedure, the saturations for each successive factor will be proportional to the larger multiplier in equations (i) or (ii), whichever of the two we have employed. They will therefore be proportional to the relative difficulty of each test, augmented if necessary to allow for chance. Thus, in the second case (where the examinees guess) let us write $f_{1i} = \frac{p_i}{\sqrt{(1 - p_i^2)}}$ and $f_{2i} = \frac{q_i}{\sqrt{(1 - p_i^2)}}$ ($i = 1, 2, \ldots n$, if there are n tests). Then the saturations will be proportional to f_{1i}; and the constant of proportionality for the ith factor will be $\sqrt{\left\{\frac{f_{2i}}{f_{1i}} - \frac{f_{2(i-1)}}{f_{1(i-1)}}\right\}}$. Now, since the triangular factors are proportional, for the two multipliers, and since the resulting correlation table will necessarily be hierarchical, it should evidently be possible (by taking each test twice according as it is the easier or harder of each pair) to represent the products or covariances as the result of a single underlying factor; and this will evidently be the factor of difficulty assumed by Mr. Gourlay at the outset.

Tables I and II have been constructed in this way from his initial figures. The unbracketed figures in Table I correspond to his Table III (items of 'same content' 'with guessing'). The analysis in Table IIA is based on the figures in Table I, *including* those in brackets, the correlations being really standardized covariations. The analysis in Table II B is based on the symmetrical correlation matrix derived from Table I by *omitting* the bracketed figures, and is given to illustrate the resulting 'triangular' factor-matrix.

TABLE I. POINT CORRELATIONS FOR HYPOTHETICAL DATA

Item	1	2	3	4	5
1	·4444	(·5601)	(·6667)	(·8729)	(1·2060)
2	·3267	·4118	(·4901)	(·6416)	(·8866)
3	·2500	·3151	·3750	(·4901)	(·6784)
4	·1455	·1833	·2182	·2857	(·3947)
5	·0335	·0422	·0502	·0656	·0909

TABLE II. FACTOR ANALYSIS

Test	Difficulty	A. Single Factor Analysis		B. Triangular Factor Analysis				
		Multipliers		Saturations				
	p	$f_1 = \frac{p}{\sqrt{(1 - p^2)}}$	$f_2 = \frac{q}{\sqrt{(1 - p^2)}}$	$\cdot 5 f_1$	$\cdot 4226 f_1$	$\cdot 4880 f_1$	$\cdot 9129 f_1$	$2 \cdot 7386 f_1$
1	·8	1·333333	·333333	·6666	·0000	·0000	·0000	·0000
2	·7	·980196	·420084	·4901	·4142	·0000	·0000	·0000
3	·6	·750000	·500000	·3750	·3169	·3660	·0000	·0000
4	·4	·436436	·654654	·2182	·1844	·2130	·3984	·0000
5	·1	·100504	·904534	·0502	·0425	·0490	·0918	·2752

However, in dealing with empirical data it would seem best to eliminate the effects of guessing before calculating and factorizing the correlations. Whether we believe (as Mr. Gourlay assumes) that the probability of guessing each of the alternatives can be assigned *a priori*, or whether we think it wiser to estimate those probabilities from actual data, the principles to be applied will be very similar to those adopted for the analogous problems that arise in genetics and allied fields: (see, for example, K. Mather, *Statistical Methods in Biology*, 1943, esp. chap. XII, 'Information and Estimation' and refs.).

ENID JOHN
CYRIL BURT

Vol. IV
Part II

The British Journal of Psychology,
Statistical Section

June,
1951

ANALYSIS OF VARIANCE WITH UNEQUAL NUMBERS

By A. DEANS PEGGS
Colonial Education Service, Bahamas

I. *Problem and Data.* II. *Analysis of Total Scores.* III. *Advantages and Limitations of the Procedure.* IV. *Regression on Age.* V. *Analyses of Scores for Separate Tests.* VI. *Summary.*

I. PROBLEM AND DATA

In planning psychological and educational researches, as a reference to actual experiments in schools will show (7), considerable practical difficulties arise when the investigator attempts to keep the numbers of his subjects (children or the like) equal in every subclass. That perhaps is one reason why analysis of variance has been, at any rate until quite recently, so seldom used by psychologists. It is the purpose of this article to describe a method of analysing variance in the case where the numbers in the subclasses are unequal, and to illustrate its value by a concrete example.[1]

TABLE I. NUMBERS OF CHILDREN, TOTAL SCORES, AND AGES

SOCIAL STATUS		COLOUR			Total
		Black	Brown	White	
High	n	5	7	17	29
	Σx	197·5	252·0	714·0	1,163·5
	Σa	106	123	254	483
Middle	n	10	20	8	38
	Σx	308·5	741·0	272·0	1,321·5
	Σa	172	389	170	731
Low	n	34	31	27	92
	Σx	1,064·0	1,057·5	964·5	3,086·0
	Σa	636	544	475	1,655
Total	n	49	58	52	159
	Σx	1,570·0	2,050·5	1,950·5	5,571·0
	Σa	914	1,056	899	2,869

$\Sigma x^2 = 216{,}842$; $\Sigma a^2 = 67{,}063$; $\Sigma ax = 107{,}525$.

The aim of the original investigation was to study the effects of race and social status on the test performances of school children in a certain colony. Only a part of the enquiry will be reported here ; and, in deference to the wishes of the investigators concerned, the precise locality will not be mentioned. The group-tests used were the following :

1. *Goodenough Man Test.* The child is instructed to draw a man, and points are awarded for the amount of detail given.
2. *Purse Test.* From Form L of the Terman-Merrill Revision, scored on a 10-point scale.
3. *Squares Test.* Similar to Koh's Blocks, but modified for use as a paper-and-pencil group-test.

[1] For advice and guidance in regard to the analysis reported here the writer is greatly indebted to Dr. D. N. Lawley, and for references and other comments to Professor Burt.

4. *Blocks Test.* The child has to estimate the number of blocks in an illustration.
5. *Forms Test.* A square has to be dissected by pencil lines to yield parts of given shape and size.
6. *Pictorial Classification.* The dissimilar figure in a series of drawings has to be crossed out.

In addition, the data include observations on (1) age of child ; (2) social status of his parents (three grades) ; (3) his race or colour (Black, Brown, White). Table I (on the preceding page) summarizes the information concerning number of children (n), total age measured from $7\frac{1}{2}$ as zero (Σa), and the total of their raw unweighted scores in all six tests (Σx), classified according to social status (parental occupation) and race or colour.

II. ANALYSIS OF TOTAL SCORES

As will be seen, the numbers in the nine subclasses are neither equal nor proportional. This means that a comparison of the score totals for the three occupations, taken as they stand, will be markedly biased by differences of score among the races or colours, if any such differences do in fact exist, and vice versa. Similarly, the distribution of age is not the same for different occupations and different races. Snedecor (2) claims that in such cases his ' Method of expected numbers ' may be used, provided the departure of the subclass numbers from proportionality is no greater than could be ascribed to chance sampling. Here, however, χ^2 is 15·08 and P lies between 0·01 and 0·001. Accordingly the method proposed instead is that used by Yates (3, 5) with similar data from experiments on livestock and plants.[1]

In the present investigation the method involves fitting constants representing the main effects of occupation and race or colour, namely, γ, representing the hypothetical mean score of Black children in the low occupation category, and α_1, α_2, β_1, β_2, being increments of score, added to this basic mean, due to the effects of Brown colour, White colour, Middle occupation, High occupation.

Table I must accordingly be modified to show, not the observed total score in each subclass, but the hypothetical subclass mean and the marginal or class totals to which they must be made to conform as nearly as possible. Age, for the moment, is omitted.

TABLE II. HYPOTHETICAL SUBCLASS MEANS ($\bar{x}$)

SOCIAL STATUS		COLOUR			Total	
		Black	Brown	White		
High ..	$\bar{x}$	$\gamma + \beta_2$	$\gamma + \alpha_1 + \beta_2$	$\gamma + \alpha_2 + \beta_2$	Σx	1,163·5
Middle ..	$\bar{x}$	$\gamma + \beta_1$	$\gamma + \alpha_1 + \beta_1$	$\gamma + \alpha_2 + \beta_1$	Σx	1,321·5
Low ..	$\bar{x}$	γ	$\gamma + \alpha_1$	$\gamma + \alpha_2$	Σx	3,086·0
Total ..	Σx	1,570·0	2,050·5	1,950·5	Σx	5,571·0

[1] The earliest discussion of the problem appears to be that of Brandt (4). In the course of his studies on sequential analysis Professor Wald encountered the problem of unequal class frequencies : his method, however, was simply to pool all sources of interaction with the error (6). Both Wilks (*Metron*, VIII, 1938, pp. 141f.) and Nair (*Sankya*, V, 1941, pp. 317f.) have touched on the problem, reaching roughly the same results as Yates. But the fullest discussion of the problem is to be found in Dr. Fei Tsao's paper on a ' General solution of the analysis of variance and covariance in the case of unequal numbers ' (9) : she points out some of the shortcomings of Yates's procedure and illustrates her discussion by referring to a study (jointly with Professor Palmer O. Johnson) in educational development (10). (I am indebted to Professor Burt for these additional references : he adds that " Tsao's paper is of special interest because the writer has educational and psychological problems in mind. Moreover, she appears to define with exceptional clearness the restrictions involved in the mathematical solution and supplements her formal solution with new approximate working methods." He also points out that those readers who do not possess easy access to the paper of Yates referred to in the text will find an excellent discussion of the problem and the procedure here described in the recent volume of M. G. Kendall's *Advanced Theory of Statistics*, Vol. II, 1946 (8, pp. 220–8), where Yates's worked example is set out in full.)

This hypothesis assumes that the effects of parental occupation and colour are additive. Departures from this assumption are termed interactions ; and, before analysing the variance for colour and occupation, it will be necessary to ascertain whether the interactions are negligible. We must first estimate the five constants. This is done by the method of least squares, and consists in minimizing the quantity $\Sigma n \times$ (observed mean − hypothetical mean)^2, where n is the number in any one subclass and the summation is over the nine subclasses. For the five unknowns this gives five equations, which may be deduced from Table II, taking the subclass frequencies from Table I. For instance, the equation (1) for α_1 is obtained from the column headed 'Brown' by equating the total hypothetical score to the observed total, thus :

$$7(\gamma + \alpha_1 + \beta_2) + 20(\gamma + \alpha_1 + \beta_1) + 31(\gamma + \alpha_1) = 2{,}050{\cdot}5.$$

The equations (2), (3), and (4) for α_2, β_1, and β_2 are similarly derived from the column headed 'White' and the rows marked 'Middle' and 'High' respectively. The fifth equation, for γ, may be obtained from the marginal (or class) totals and the grand total. Collecting terms, we have :

$$58\alpha_1 + 20\beta_1 + 7\beta_2 + 58\gamma = 2{,}050{\cdot}5 \quad (1)$$

$$52\alpha_2 + 8\beta_1 + 17\beta_2 + 52\gamma = 1{,}950{\cdot}5 \quad (2)$$

$$20\alpha_1 + 8\alpha_2 + 38\beta_1 + 38\gamma = 1{,}321{\cdot}5 \quad (3)$$

$$7\alpha_1 + 17\alpha_2 + 29\beta_2 + 29\gamma = 1{,}163{\cdot}5 \quad (4)$$

$$58\alpha_1 + 52\alpha_2 + 38\beta_1 + 29\beta_2 + 159\gamma = 5{,}571{\cdot}0 \quad (5)$$

Writing these equations in the form $A\lambda = s$, we have $\lambda = A^{-1}s$. An approximation to the reciprocal, A^{-1}, was first obtained by pivotal condensation and then corrected to 8 decimal places by Hotelling's iterative method (1). From this we obtain the following estimates for the vector λ :

$$\alpha_1 = 3{\cdot}0663,\ \alpha_2 = 4{\cdot}2543,\ \beta_1 = 1{\cdot}0051,\ \beta_2 = 5{\cdot}6249,\ \gamma = 31{\cdot}2617.$$

It is now possible to rewrite Table II, substituting these values for the five constants ; and then, by multiplying the hypothetical subclass means so reached by the number of children in each case, to obtain the hypothetical total score for each subclass. These are shown in Table III.

TABLE III. ESTIMATED HYPOTHETICAL SUBCLASS MEANS

Social Status		Colour			Total
		Black	Brown	White	
High ..	$\bar{x}$	36·8866	39·9529	41·1409	—
	Σx	184·4330	279·6703	699·3953	1,163·4986
Middle ..	$\bar{x}$	32·2668	35·3331	36·5211	—
	Σx	322·6680	706·6620	292·1688	1,321·4988
Low ..	$\bar{x}$	31·2617	34·3280	35·5160	—
	Σx	1,062·8978	1,064·1680	958·9320	3,085·9978
Total ..	Σx	1,569·9988	2,050·5003	1,950·4961	5,570·9952

The significance of the interaction can now be tested. The sum of squares between the nine subclasses, as calculated either from the observed data of Table I or from the hypothetical values of Table III, prove to be 1,711 and 1,422 respectively. The sum of squares for interaction is the difference between these two, i.e., 289. In practice it is unnecessary to construct Table III ; the sum of squares, 1,422, may be derived by forming the inner product of the constants (λ) and the totals (s), from which the correction for the mean, 195,195, may then be subtracted. Table IV shows the test for the significance of the interaction (second line) :

TABLE IV. TEST OF SIGNIFICANCE OF INTERACTION AND OF COMBINED MAIN EFFECTS

Source of Variation	Degrees of Freedom	Sum of Squares	Mean-square	Variance Ratio	Significance
Colour (C) and Occupation (O)	4	1,422	356	2·68	S
Interaction (C × O)	4	289	72	0·54	NS
Between subclasses	8	1,711	—	—	—
Within subclasses	150	19,936	132·91	—	—
Total	158	21,647	—	—	—

NS = Not significant at 5 per cent. level ; S, SS, and SSS = Significant at 5 per cent., 1 per cent., and 0·1 per cent. levels respectively.

There is no evidence of significant interaction. We may therefore assume that the effects of colour and occupation are additive. We can now test the significance of the main effects. The figures for this are also shown in Table IV. With 4 degrees of freedom the theoretical variance ratio at the 5 per cent. level is 2·43. The combined effect of occupation and race is therefore significant. It would have been permissible to combine both the degrees of freedom and the sum of squares for 'interaction' with those for 'within subclasses' (so-called 'error'); but, since this procedure would here make no considerable difference, it has not been carried out.

To test the significance of occupation we need an estimate of variance which is completely free from the influence of colour. On account of the unequal numbers, the quantity

$$\frac{(1{,}163{\cdot}5)^2}{29} + \frac{(1{,}321{\cdot}5)^2}{38} + \frac{(3{,}086{\cdot}0)^2}{92} - \frac{(5{,}571{\cdot}0)^2}{159} = 957$$

(which would ordinarily be used) must here be partly due to race. This can be overcome by subtracting from 1,422 (the sum of squares due to the combined effect of race and occupation) the quantity

$$\frac{(1{,}570{\cdot}0)^2}{49} + \frac{(2{,}050{\cdot}5)^2}{58} + \frac{(1{,}950{\cdot}5)^2}{52} - \frac{(5{,}571{\cdot}0)^2}{159} = 764,$$

which includes all the effect of colour as well as some due to occupation. The resulting estimate, 658, probably does not represent the full effect of occupation, but is due to nothing else but occupation. A similar procedure can be adopted to obtain an unbiased estimate for variance due to colour alone. To decide the significance the variance is then compared with the figure previously found, viz., 132·91 (Table V).

TABLE V. TEST OF SIGNIFICANCE OF MAIN EFFECTS

Source of Variation	Degrees of Freedom	Sum of Squares	Mean-square	Variance Ratio	Significance
Colour and occupation	4	1,422	—	—	—
Colour, neglecting occupation	2	764	—	—	—
Remainder (occupation) ..	2	658	329	2·48	NS
Colour and occupation	4	1,422	—	—	—
Occupation, neglecting colour ..	2	957	—	—	—
Remainder (colour)	2	465	233	1·75	NS

The analysis shows that the effects of occupation and colour, considered separately, are not significant. Since the previous analysis has shown that the two together do produce a significant effect, more extensive data would be required to assess the magnitude of each factor separately.

III. ADVANTAGES AND LIMITATIONS OF THE PROCEDURE

From the foregoing illustration some of the merits of the method proposed will now be obvious. In his discussion of ' The applicability of analysis of variance to educational problems ' (7), Burt has summarized the chief advantages of analysing variance as a research procedure in psychology. He emphasizes the fact that, with its aid, " we can allow more than one factor to vary concurrently " and still apportion the relative influence of the several causes ; as a consequence, " the conditions of the experiment become more natural," and we can thus avoid the " artificial restrictions introduced by experimental control." [1] In particular, there is " then no need for a separate control group " ; and as a result, the enquiry can be carried out in circumstances which approach those of the children's ordinary activities. Two special cautions, however, must be stressed. The freedom allowed does not absolve the investigator from the responsibility of carefully planning the experiment beforehand. And there is the further danger that results obtained may in part be due to unsuspected factors other than those investigated : for instance, in the present study, sex, school, or age. Accordingly, in any conclusion drawn, room must always be allowed for alternative explanations of the phenomena observed.

Here, in point of fact, the influence of sex and school proved to be non-significant, but that of age was fully significant. Consequently, in virtue of the disproportion in numbers, the effects of age tend to obscure those of race and occupation. We must therefore endeavour to remove this influence.

IV. REGRESSION ON AGE

To test the significance of the regression of score on age we must know the ' within classes ' (or ' error ') sum of squares for age and the corresponding sum of cross-products for age and score. They are shown at the foot of Table I as 67,063 and 107,525 respectively. From these is obtained the estimate, $\frac{(7,315)^2}{14,923} = 3,586$, of that part of the ' error ' sum of squares which is due to the regression of score on age. The analysis appears in Table VI.

TABLE VI. TEST OF SIGNIFICANCE OF REGRESSION OF SCORE ON AGE

Source of Variation	Degrees of Freedom	Sum of Squares	Mean-square	Variance Ratio	Significance
Linear regression	1	3,586	3,586	32·68	SSS
Remainder	149	16,350	109·73	—	—
Within subclasses	150	19,936	—	—	—

[1] Burt, C., *Brit. J. Educ. Psychol.*, XVI, pp. 129–32. It should be noted that in a symposium on ' Recent developments of statistical method in psychology ' (*Proc. Roy. Soc.*, B, CXXV, 1938, pp. 417f.) several statistical psychologists, including both Spearman and W. Stephenson, opposed the notion that the newer methods such as analysis of variance and maximum likelihood might (as suggested by Burt) be applied to problems in psychological and educational research. Most of their objections seem since to have been fully met. Nevertheless, their criticisms serve to bring out the limitations of the method and the need for carefully *planning* experiments in which such methods are to be applied : see, for example, Edwards, A. L., *Experimental Design in Psychological Research*, 1950, especially chap. X and following.

That the regression of score on age is highly significant was to be expected. It indicates that the removal of age will increase the significance of the main effects, since the error mean-square is now reduced from 132·91 to 109·73.

Constants can be fitted for age from the equation $\lambda' = A^{-1}s_1$, where s_1 is obtained from the marginal totals for age in the last column of Table I: i.e.,

$$s_1 = (1{,}056,\ 899,\ 731,\ 483,\ 2{,}869)$$

We obtain $\lambda' = (-0{\cdot}6054,\ -1{\cdot}0572,\ 1{\cdot}2746,\ -1{\cdot}0824,\ 18{\cdot}5034)$. From these values Table VII may be constructed showing the hypothetical subclass mean ages on the assumption of negligible interaction.

TABLE VII. HYPOTHETICAL SUBCLASS MEAN AND TOTAL AGES

SOCIAL STATUS		RACE			Total
		Black	Brown	White	
High ..	$\bar{a}$	17·4210	16·8156	16·3638	—
	Σa	87·1050	117·7092	278·1846	482·9988
Middle ..	a	19·7780	19·1726	18·7208	—
	Σa	197·7800	383·4520	149·7664	730·9984
Low ..	$\bar{a}$	18·5034	17·8980	17·4462	—
	Σa	629·1156	554·8380	471·0474	1,655·0010
Total ..	Σa	914·0006	1,055·9992	898·9984	2,868·9982

TABLE VIII. COMPLETE TABLE OF SUMS OF SQUARES AND CROSS-PRODUCTS OF SCORE ON AGE

	Source of Variation	1 d.f.	2 Σa^2	3 Σax	4 Σx^2	5 $\frac{(\Sigma ax)^2}{\Sigma a^2}$	6 (4)–(5)	7 d.f.
(a)	Subclasses and mean ..	9	52,140	100,210	196,906	—	—	—
(b)	Colour, occupation and mean	5	51,905	100,204	196,617	—	—	—
(c)	Colour and mean (neglecting occupation)	3	51,818	100,340	195,959	—	—	—
(d)	Occupation and mean (neglecting colour)	3	51,879	100,314	196,152	—	—	—
(e)	Mean	1	51,768	100,523	195,195	—	—	—
(f)	Within classes (error) ..	150	14,923	7,315	19,936	3,586	16,350	149
(A)	(a)−(b)+(f)	154	15,158	7,321	20,225	3,536	16,689	153
(B)	(b)−(e)+(f)	154	15,060	6,996	21,358	3,250	18,108	153
(C)	(b)−(d)+(f)	152	14,949	7,205	20,401	3,473	16,928	151
(D)	(b)−(c)+(f)	152	15,010	7,179	20,594	3,434	17,160	151
(i)	(a)−(b) Interaction ..	4	235	6	289	(A)−(f)	339	4
(ii)	(b)−(e) Colour and occupation	4	137	−319	1,422	(B)−(f)	1,758	4
(iii)	(b)−(d) Colour	2	26	−110	465	(C)−(f)	578	2
(iv)	(b)−(c) Occupation ..	2	87	−136	658	(D)−(f)	810	2

From the values there given the hypothetical sum of squares of age between subclasses is found to be 137. But, as before, it may be derived directly from the inner product of λ' and s_1. The corresponding sum of cross-products (−319) may be obtained either from Tables III and VII or from the inner product of λ and s_1 or of λ' and s.

The remaining sums of squares of age and sums of cross-products needed for removing the effect of age are obtained in the usual way (Table VIII). Note that the first four rows are not corrected for the mean ; this is accomplished incidentally in the subtractions at (A) to (D).

The quantities in columns 2 to 4 at (A) to (D) correspond to those at (i) to (iv), but include the error sum of squares or cross-products (f). The sums of squares of scores at (A) to (D) in column 4 are adjusted by subtracting from each the quantity $\frac{(\text{sum of cross-products})^2}{\text{sum of squares for age}}$ (column 5). From the corresponding remainders in column 6 is subtracted the adjusted error sum of squares, 16,350. These final adjusted sums of squares for use in the final analysis of variance are the last four entries in column 6. Thus Table VIII contains all the information necessary for the analyses made in Tables IV, V, VI, and IX.

The mean-squares corresponding to the reduced sums of squares are now tested for significance against the reduced error mean-square, 109·73. This is shown in Table IX. The effect of race (or ' colour ') and social status (or ' occupation ') together is now significant at 1 per cent. Occupation, tested separately, can be adjudged significant at the 5 per cent. level ; but there is still no justification for attributing any significant influence to colour.

TABLE IX. FINAL TEST OF SIGNIFICANCE : EFFECT OF AGE REMOVED

Source of Variation	Degrees of Freedom	Sum of Squares	Mean-square	Variance Ratio	Significance
Interaction (C × O)	4	339	85	0·77	NS
Colour and occupation	4	1,758	440	4·01	SS
Colour	2	578	289	2·63	NS
Occupation	2	810	405	3·69	S
Error	149	16,350	109·73	—	—

V. ANALYSES OF SCORES FOR SEPARATE TESTS

As explained above, the figures analysed in Table I were the simple sums of the children's unweighted scores in each of the six tests. From the figures given in Table X it will be seen that standard deviations of the scores showed considerable variation. Hence the contribution of some of the tests (tests 1 and 4, for instance) was much greater than that of the others. Since there were no *a priori* grounds for weighting the six tests in any specific ratio, the procedure outlined was applied to each of the tests separately. The results, with age removed, are summarized in Table X. In no instance was there found to be significant interaction.

TABLE X. SIGNIFICANCE OF MAIN EFFECTS AS REVEALED BY INDIVIDUAL TESTS

Test	1 Man	2 Purse	3 Squares	4 Blocks	5 Forms	6 Classification
Standard deviation	5·9	2·7	1·8	3·0	1·2	2·8
Regression of score on age ..	SSS	SS	SSS	SSS	SS	S
Colour and occupation ..	NS	NS	S	SS	SSS	SSS
Colour	NS	NS	NS	NS	SSS	SSS
Occupation	NS	NS	S	S	SSS	SSS

VI. SUMMARY

1. In a research intended to investigate the effect of race and social status on cognitive tests, the numbers of pupils involved in the two-way classification were found to be disproportionate or ' non-orthogonal.' A chi-squared test showed that the deviations from proportionality could not be ascribed to chance. Hence Snedecor's ' Method of expected numbers ' was not applicable. Instead Yates's method was adopted for fitting constants for the main effects and for the mean by the method of least squares.

2. The interactions between the main effects were found to be non-significant ; and it was consequently inferred that these effects may legitimately be treated as additive. It is then shown how an unbiased estimate of the variance resulting from the several main effects may be secured.

3. The regression of score on age was found to be fully significant. By the method of covariance, involving the fitting of a further constant for age, the effect of age was removed. The significance of the main effects was tested afresh, and found to be enhanced. This process was repeated for each test separately ; and the significance of the main effects was found to vary considerably from test to test.

REFERENCES

1. Hotelling, H. (1943). 'Some new methods in matrix calculation,' *Ann. Math. Stat.*, XIV, 1–43.
2. Snedecor, G. W. (1934). ' The method of expected numbers for tables of multiple classification with disproportionate subclass numbers,' *J. Amer. Stat. Assoc.*, XXIX, 389–93.
3. Yates, F. (1933). ' The principles of orthogonality and confounding in replicated experiments,' *J. Agric. Sc.*, XXIII, 108–45.
4. Brandt, A. E. (1933). ' The analysis of variance in $2 \times s$ tables with disproportionate frequencies,' *J. Amer. Stat. Assoc.*, XXVIII, 164–73.
5. Yates, F. (1934). ' Analysis of multiple classifications with unequal numbers in different classes,' *J. Amer. Stat. Assoc.*, XXIX, 51–66.
6. Wald, A. (1941). ' On the analysis of variance in the case of multiple classifications with unequal class frequencies,' *Ann. Math. Stat.*, XII, 346–50.
7. Burt, C. (1946). ' The applicability of analysis of variance to educational problems,' *Brit. J. Educ. Psychol.*, XVI, 129–32.
8. Kendall, M. G. (1946). *The Advanced Theory of Statistics*, Vol. II. London : C. Griffin & Co.
9. Tsao, F. (1946). ' General solution of the analysis of variance and covariance in the case of unequal or disproportionate numbers of observations in the subclasses,' *Psychometrika*, XI, 107–28.
10. Johnson, Palmer O. (1949). *Statistical Methods in Research.* New York : Prentice-Hall.

Vol. IV
Part II

The British Journal of Psychology,
Statistical Section

June,
1951

INCOMPLETE BLOCKS IN RANKING EXPERIMENTS

By J. DURBIN
London School of Economics

I. *Introduction.* II. *The Coefficient of Concordance.* III. *Test of Independence.* IV. *A Numerical Example.*

I. INTRODUCTION

Problem. Suppose we have n objects, and wish to investigate the preferences shown by m different observers : one method will be to ask each observer to rank the objects in order of preference. This is the ordinary case of m rankings described by Kendall (1, p. 80). If, however, n is large, the method is often unsatisfactory, since it is difficult to rank a large number. Further, if the objects are presented as a list of names, the order in which they are presented may become a disturbing factor as n increases. In such cases a better method is to present them in small blocks. In this paper the use of ' balanced incomplete blocks ' will be examined, and expressions calculated for the mean and variance of the coefficient of concordance. These will enable approximate tests of independence among the rankings to be constructed. Finally, the optimal properties of the estimate of the true ranking will be discussed.

The need for some such procedure is obvious ; and the method to be described has a number of applications in social and psychological research. For instance, Hall and Caradoc Jones (3) have recently described a survey in which 1,399 people were asked to grade 30 occupations in order of social standing. The difficulty of grading as many as 30 items was admitted at the outset ; and it was therefore suggested that the respondents should first assign each occupation to one of five groups, and then order the members of each group separately. This, however, offers only a partial relief ; and in many similar enquiries it would seem to be preferable for the investigator to arrange the objects in smaller groups beforehand. Much the same device may also be needed in studying preferences and in experiments on discrimination.

Balanced Incomplete Blocks. To be workable in practice, a ranking layout which uses incomplete blocks should yield results that permit of a fairly straightforward analysis. This requirement gives rise to the following two conditions, which seem requisite in themselves on grounds of common sense :

(*a*) Each object should occur an equal number of times in the experiment as a whole.
(*b*) The number of times two particular objects occur together in the same block should be the same for all possible pairs of objects.

Designs in which these conditions are fulfilled are called ' balanced incomplete block designs.' An excellent account of them, together with a detailed list of the appropriate layouts, will be found in the recent book by Cochran and Cox (2).

As stated above, it will sometimes be desirable to eliminate from the comparisons the effect of the order of presentation within the block. This can be done by arranging the layout so that every object occurs equally often in each position in a block. Such an arrangement can always be found by taking a sufficient number of blocks to cover all the requisite combinations ; but this number will frequently be impracticably large. However, certain compact designs satisfying the condition can be arranged for multiples of $p^2 + p + 1$ objects in blocks of $p + 1$; these are called ' Youden square designs,' and are fully described by Cochran and Cox.

It is unnecessary for each observer to rank the objects in *all* the blocks in a single repetition. For instance, an observer may be called upon to rank only the objects in a single block. The tests of independence described below apply to this case equally with that in which he is required to rank all the n objects. If, however, the purpose is to estimate the mean ranking for a population of observers, it may be desirable to ask the observers to rank the objects more than once, using

different arrangements in the blocks. In this way a greater number of independent comparisons between pairs can be obtained. With such an arrangement the greatest precision is achieved when each observer ranks an equal number of blocks. There will then be no question of testing for independence among the different rankings, except when the hypothesis under consideration is that each observer allocates his preferences at random in successive rankings.

A Simple Example. Suppose it is desired to rank seven objects A B C D E F G in blocks of three. From the list of Youden square designs in Cochran and Cox we take the following design :

A	B	C	D	E	F	G
B	C	D	E	F	G	A
D	E	F	G	A	B	C

Each column represents a block, and each row a single replication of the objects. Since only three objects out of seven appear in each column, the ' blocks ' are ' incomplete ' ; and, since the arrangement fulfils conditions *a* and *b*, it may be described as ' balanced.' By permuting the letters A, B, . . ., or by repeating the design as it stands, we can obtain further sets of seven blocks, if we wish. The number of blocks assigned to each observer is irrelevant to the problem of testing independence ; and in any practical case would depend on the purpose of the experiment.

Suppose the following ranks were allotted :

2	2	1	1	2	2	3
1	3	2	3	3	3	1
3	1	3	2	1	1	2

The total ranks for the seven objects will be as follows :

A	B	C	D	E	F	G
4	4	6	6	6	8	8

If the ranks had been allotted at random within the blocks, these totals would tend to be equal. If, on the other hand, the ranks corresponded to a strict ordering of the objects, the totals would be 3, 4, 5, 6, 7, 8, 9, and would differ among themselves as widely as possible. Thus, the variance of the total ranks should be a good statistic for testing the independence of the rankings within the different blocks. It will be defined more precisely in the following section.

II. THE COEFFICIENT OF CONCORDANCE

Mean and Variance of the Coefficient. Suppose the n objects are presented in blocks of k, and that in the experiment as a whole each is ranked m times. Then the number of blocks will be $\frac{mn}{k}$. Within each block there are $\frac{1}{2}k(k-1)$ comparisons between pairs. Consequently, the total number of comparisons is $\frac{1}{2}mn(k-1)$. Let λ be the number of blocks in which a particular pair of objects occurs. Then $\frac{1}{2}n(n-1)\lambda = \frac{1}{2}mn(k-1)$ so that $\lambda = \frac{m(k-1)}{n-1}$.

Consider first the case in which there are no ties. Let x_{ij} denote the rank assigned to the jth object in the ith replication, and S the sum of squared deviations from the mean of the totals $x_j = \sum_{i=1}^{m} x_{ij}$. S attains its maximum when there is perfect concordance between the rankings. The total ranks are then m, $m + \lambda$, $m + 2\lambda$, . . . $m + (n-1)\lambda$. Thus the maximum value of S is $\frac{\lambda^2 n(n^2-1)}{12}$.

The ' coefficient of concordance ' (1, p. 81) is therefore defined by

$$W = \frac{\text{observed } S}{\text{maximum } S}$$

$$= \frac{12S}{\lambda^2 n(n^2-1)}$$

$$= \frac{12\Sigma x_j^2 - 3nm^2(k+1)^2}{\lambda^2 n(n^2-1)}. \qquad (1)$$

To calculate the mean and variance, we take the x's as measured from the block means. Following the derivation given by Kendall (*loc. cit.*, pp. 90f.), we have,

$$W = \frac{12\,\Sigma x_{ij}^2 + 24 \sum_j \sum_{i<p} x_{ij}\, x_{pj}}{\lambda^2 n\,(n^2-1)},$$

where $$\Sigma x_{ij}^2 = \frac{mn\,(k^2-1)}{12}.$$

Thus $$W = \frac{k+1}{\lambda\,(n+1)} + \frac{24\,U}{\lambda^2 n\,(n^2-1)}. \qquad (2)$$

We therefore require the mean and variance of $U = \sum_j \sum_{i<p} x_{ij}\, x_{pj}$. The jth object occurs in different blocks in the ith and pth replications ; and an x in any one block is independent of an x in any other block. Consequently,

$$E(U) = 0 \text{ ; and}$$

$$U^2 = \Sigma\, x_{ij}^2\, x_{pj}^2 + 2\,\Sigma\, x_{ij}\, x_{pj}\, x_{ql}\, x_{rl}. \qquad (3)$$

There are $\frac{1}{2}nm\,(m-1)$ terms in the first of these sums ; and the expectation of each is $\{E(x_{ij}^2)\}^2 = \frac{(k^2-1)^2}{144}$. The only terms contributing anything to the expectation of the second sum are those for which the jth and lth objects occur together in each of two different blocks. Now $E(x_{ij}\, x_{ql}) = E(x_{pj}\, x_{rl}) = -\frac{k+1}{12}$ when the jth and lth objects are in the same block, and zero when they are not in the same block. Thus there is a contribution $2\left(\frac{k+1}{12}\right)^2$ whenever the jth and lth objects occur together in each of two different blocks. The number of such cases is $\frac{1}{4}n\,(n-1)\,\lambda\,(\lambda-1)$. Substituting in the expectation of (3) we have

$$E(U^2) = \frac{1}{288}\{nm\,(m-1)\,(k^2-1)^2 + n\,(n-1)\,\lambda\,(\lambda-1)\,(k+1)^2\}$$

$$= \frac{mn\,(k+1)\,(k^2-1)}{288}\{(m-1)\,(k-1) + \lambda - 1\},$$

since $$(n-1)\lambda = m\,(k-1).$$

Consequently $$E(W) = \frac{k+1}{\lambda\,(n+1)}$$

and $$var\,(W) = \frac{2\,mn\,(k+1)\,(k^2-1)}{\lambda^4 n^2\,(n^2-1)^2}\{(m-1)\,(k-1) + \lambda - 1\}$$

$$= \frac{2\,(k+1)^2}{mn\,\lambda^2\,(n+1)^2}\left(m - 1 + \frac{\lambda-1}{k-1}\right). \qquad (4)$$

When $k = n$, and therefore $\lambda = m$ (Kendall's case of 'm rankings'), we have

$$E(W) = \frac{1}{m},$$

$$var\,(W) = \frac{2\,(m-1)}{m^3\,(n-1)},$$

which agree with the formulæ given by Kendall (*loc. cit.*, p. 93).

The relative simplicity and the generality of these formulæ are due to the symmetry of the design, and in particular to the fact that each pair of objects occurs together an equal number of times in the same block. General formulæ for the higher moments cannot be obtained, since they would necessitate sets of three, four, etc., objects occurring equaly frequently in the same block ; and these conditions will only be satisfied by certain of the less useful designs. However, the higher moments are not needed, since the approximate distributions used for testing independence in the way described below require the first two moments only.

Paired comparisons. The case of paired comparisons, described in general terms by Kendall (pp. 121f.), is of special interest. Here $k = 2$; and to give the necessary symmetry n must be even, and m must be a multiple of $n - 1$. We find

$$E(W) = \frac{3(n-1)}{m(n+1)},$$

and

$$var(W) = \frac{18(n-1)^2}{m^3 n(n+1)^2}\left(\frac{mn}{n-1} - 2\right).$$

If $m = n - 1$, then $E(W) = \frac{3}{n+1}$, and $var(W) = \frac{18(n-2)}{n(n^2-1)(n+1)}$.

Ties. The above results hold only when there are no ties in the rankings. However, for the case $k=n$, Kendall (p. 94) has shown that the presence of a small number of ties does not affect the moments of W appreciably; and the result may be expected to apply equally to the incomplete block case. The appropriate modifications for a large number of ties are given by Kendall for the case $k=n$; but for large m they are already rather complicated and for other values of k they would be still more difficult to apply. A simpler procedure is to allocate ranks of tied objects at random, e.g., by tossing a coin. This does not affect the validity of tests based on W, and enables the same formulæ to be used in all cases.

A further method consists in allocating the ranks of tied members in such a way that the significance of W is reduced. The effect of this is to make the test rather more stringent than the significance level would suggest.

III. TEST OF INDEPENDENCE

Following Kendall we note that there are two approximations to the distribution of W which are suitable for testing the hypothesis of independence among the m rankings. The more accurate of these is the Beta distribution:

$$dF = \frac{1}{B(p,q)} W^{p-1}(1-W)^{q-1}\,dW. \qquad (5)$$

Kendall (p. 93) has shown by calculating the third and fourth moments that for the case $k = n$ this gives a good approximation. By analogy one would expect the approximation to be satisfactory in the incomplete block case. Equating the first two moments with the mean and variance of W we find we must take

$$\left.\begin{aligned} p &= \frac{mn\left(1 - \frac{k+1}{\lambda(n+1)}\right)}{2\left(\frac{nm}{n-1} - \frac{k}{k-1}\right)} - \frac{k+1}{\lambda(n+1)} \\ \text{and} \quad q &= \left(\frac{\lambda(n+1)}{k+1} - 1\right)p \end{aligned}\right\}, \qquad (6)$$

where $\lambda = \frac{m(k-1)}{n-1}$ as before.

When $k = n$, these values reduce to $p = \frac{1}{2}(n-1) - \frac{1}{m}$ and $q = (m-1)p$, as given by Kendall.

On putting $F = \frac{\left(\frac{\lambda(n+1)}{k+1} - 1\right)W}{1-W}$, equation (5) reduces to Fisher's variance ratio distribution with $\nu_1 = 2p$, and $\nu_2 = 2q$ degrees of freedom, the values of p and q being given by (6).

For moderate n and large m, W tends to be distributed as a multiple of χ^2 with $n-1$ degrees of freedom, since S is a sum of squares of n variates connected by a linear relation, where the variate values are themselves sums of m independent observations.

Putting $W = a\chi^2$, we have $E(W) = \frac{k+1}{\lambda(n+1)} = a(n-1)$. Thus $\chi_r^2 = \frac{\lambda(n^2-1)}{k+1}W$ tends to be distributed as χ^2 with $n-1$ degrees of freedom.

It is of interest to compare the variance of χ_r^2 with $2(n-1)$, the variance of the fitted distribution.

$$var\,(\chi_r^2) = \frac{\lambda^2(n^2-1)^2}{(k+1)^2}\,\frac{2(k+1)^2}{\lambda^2 mn(n+1)^2}\left(\frac{mn}{n-1}-\frac{k}{k-1}\right)$$
$$= 2(n-1)\left\{1-\frac{k(n-1)}{mn(k-1)}\right\}$$
$$= 2(n-1)\left(1-\frac{1}{m}\right) \text{ approximately.}$$

(In the ordinary case of m rankings $k = n$, so that $var\,(\chi_r^2)$ equals this value exactly.) This illustrates the fact that the governing factor in the accuracy of the χ^2 approximation is the magnitude of m.

IV. A NUMERICAL EXAMPLE

Consider by way of illustration the following example with imaginary data. Using the design already given (p. 86), a manufacturer has tested the preferences of a number of persons for seven varieties of ice-cream. The letters will now refer to the seven varieties; and, as before, each column will represent a block. The design was repeated three times; and the resulting 21 blocks were assigned at random to 21 tasters, each of whom was asked to rank the three relevant varieties in order of preference. The totals of the ranks assigned to the seven objects were (let us suppose) as follows:

A	B	C	D	E	F	G
20	13	18	25	22	12	16

We require to test whether the preferences exhibit a significant departure from randomness.

The mean of the rank totals is 18. And $S = 2^2 + 5^2 + 0^2 + 7^2 + 4^2 + 6^2 + 2^2 = 134$. Further, $n = 7$, $m = 9$, and $k = 3$: so that $\lambda = \frac{9 \times 2}{6} = 3$. Hence,

$$W = \frac{12\,S}{\lambda^2 n(n^2-1)} = \cdot 532.$$

To test whether this value is significant we apply formula (6). We obtain $p = 2{\cdot}75$, and $q = 13{\cdot}75$. To test significance we require the 5 per cent. value of F for $\nu_1 = 5{\cdot}5$ and $\nu_2 = 27{\cdot}5$ degrees of freedom. This is not tabulated; but for $\nu_1 = 5$ and $\nu_2 = 27$ we have $F = 2{\cdot}57$. The corresponding value of W is given by $2{\cdot}57 = \left(\frac{3 \times 8}{4} - 1\right) W/(1-W)$, so that W = ·339. The observed value is greater than this; and therefore must certainly be greater than the value obtained from F for $\nu_1 = 5{\cdot}5$ and $\nu_2 = 27{\cdot}5$. We infer that the observed value of W is significant at the 5 per cent. level, and reject the hypothesis that preferences were random.

Estimate of a true ranking. A basic property of Kendall's coefficient τ is that it is simply related to the number of inversions of the natural order. Thus if the order of the objects is arranged so that one of the rankings is 1, 2, . . ., n, then $\tau = 1 - \frac{2Q}{\frac{1}{2}n(n-1)}$, where Q is the number of inversions in the second set of ranks. For example, in the rankings

1	2	3	4	5
2	4	1	3	5,

$Q = 3$, one for each of the inversions 21, 41, 43.

A similar property holds for the rank correlation, namely, $\rho = 1 - \frac{12\,V}{n^3-n}$, where V is the sum of the weighted inversions; here V is similar to Q except that each inversion is weighted by the difference between the two ranks in one of the rankings. For instance, in the above example we score 1 for the inversion 21, 3 for the inversion 41, and 1 for the inversion 43, giving $V = 5$. This gives $\rho = 1 - \frac{6 \times 10}{n^3-n}$, which agrees with the usual formula $\rho = 1 - \frac{6\Sigma d^2}{n^3-n}$, since $\Sigma d^2 = 10$.

The obvious way of estimating the true ranking in a balanced design is to rank the objects according to the sums of the ranks allotted to them by the different observers. Kendall (p. 97) has shown for the case $k = n$ that this method of estimation maximizes the average value of ρ for the correlation between the estimated and the observed rankings. It follows from the preceding paragraph that it also minimizes the sum of the weighted inversions in the observed rankings relative to the estimated ranking. This conclusion evidently remains true for the more general case of $k<n$.

Similar results hold for the case of tied ranks. If two objects are ranked equally in one ranking and strictly ordered in another, we call the comparison a violated tie and assign a score equal to half the corresponding score for an inversion. If the objects are tied in both rankings, there is, of course, no score. With these rules V becomes the sum of the weighted inversions plus half the sum of the weighted violated ties. Thus, if the true ranking is estimated by ranking according to the sums of ranks, where ties have been dealt with by allotting fractional ranks, the method of estimation is such as to minimize the sum of the weighted inversions in the observed rankings relative to the estimated rankings, plus half the sum of the weighted violated ties.

REFERENCES

1. Kendall, M. G. (1948). *Rank Correlation Methods.* London : Griffin.
2. Cochran, W. G., and Cox, G. M. (1950). *Experimental Designs.* New York : Wiley.
3. Hall, T., and Jones, D. Caradoc (1950). *Brit. J. Sociol.*, I., 31.

Vol. IV
Part II

The British Journal of Psychology,
Statistical Section

June,
1951

AN ITEM-ANALYSIS OF THE PROGRESSIVE MATRICES TEST

By CHARLOTTE BANKS and UMA SINHA
University College, London

I. *Procedure.* II. *Results.* III. *Summary and Conclusions.*

I. PROCEDURE

Problem. The so-called ' matrix test,' an extended version of the non-verbal analogies test,[1] was originally devised as a supplementary test for dull and backward school children(1, p. 234), and was later developed by Raven for use with mentally defective adults (2). In 1940, Burt, as a member of a Psychological Committee appointed by the Adjutant-General, suggested it for tentative inclusion in the batteries of tests to be used with Army recruits, in order to compensate for the educational bias of group-tests of the more familiar verbal type. As a result, a large amount of data on the test is now available for adults.

Vernon and Parry, in surveying the experience thus gained, report that the value of the test fell rather below expectation ; they quote the conclusion reached by Fraser-Roberts, who " was able to show, by detailed comparison with more reliable tests, that its unreliability is greatest in the 15–30 score range," i.e., at the most crucial level for classificatory purposes. As one of the possible explanations, they point out that several factorial analyses have shown that, besides the general cognitive factor of ' intelligence,' the test also involves a more specialized perceptual or visuo-spatial factor : this observation is fully in keeping with the results obtained by one of the present writers (5, p. 87). The irrelevant factor could easily be neutralized if the test were used in combination with others ; and perhaps the chief criticism of the claims commonly made on its behalf is that it is so frequently put forward as a single, self-sufficient test of intelligence. Such a claim no doubt accounts largely for its popularity ; but it is in sharp contradiction to the well-established principle that any effective test, or battery of tests, should be of a heterogeneous nature, in order to neutralize the effects of the inevitable specific factors. Thus, in the Services, as might indeed have been anticipated, wherever a composite battery of verbal and non-verbal tests was employed, " the combined result yielded a considerably better measure than did the matrices test alone " (7, p. 235).

More recently, however, Raven himself has reported highly satisfactory results from its use with children (4) ; and it has been frequently proposed as a test for sorting out school pupils for allocation to secondary schools of various types. The object of the present research was therefore to investigate a little more fully than has hitherto been done its apparent value as an intelligence test for children of school age. No previous investigator appears to have attempted what after all is the first essential in any such test—a comparative validation of the separate items ; nor has

[1] Several recent writers, including Raven himself, apparently suppose that the analogies test was devised by Spearman " to illustrate the fundamental principles of educing relations and then educing correlates." Spearman, however, explicitly states that it was ' invented by Burt ' (*Abilities of Man*, 1926, p. 201) ; and it was the high correlations obtained with this type of test, as contrasted with tests of sensory discrimination, that led Spearman to abandon his original hypothesis that ' general ability ' is identifiable with ' general sensory discrimination,' and accept the view that " *g* essentially involves the apprehension and application of logical relations " (*loc. cit.*, pp. 165f., 411). It is, as our results clearly show, precisely where this double process comes into play that matrix-items are most effective.

any writer examined their logical structure to see how far each problem really involves the particular mental processes which the test is supposed to elicit.

Subjects and Tests. The children tested consisted of 140 boys and 170 girls, i.e., 310 children in all, ranging in age from 8 to 13. The older children, viz., those over 11 years, were drawn from three secondary modern schools in the East End of London. The younger children were drawn from a primary school in the same district. The 1938 version of Raven's ' Progressive matrices ' (3) was applied to all the groups. The older children were also tested for intelligence with Richardson's Simplex test ; the younger children were tested individually with the Terman-Merrill revision of the Binet scale.[1]

Figures have been calculated for the several age-groups separately ; and, except where otherwise stated, the values cited are averages of the figures so obtained.

II. RESULTS

(A) *Reliability.* Three methods of assessing reliability were employed. With the split half method (duly corrected) the average reliability coefficient was 0·86 ; with Burt's analysis of variance method (8, p. 87), 0·90 ; and with the Kuder-Richardson formula, 0·91. The average of these figures (0·88) is appreciably smaller than that given by Raven (who used the test-retest method with a much smaller group), and barely reaches the minimum usually thought necessary for a satisfactory intelligence test.

(B) *Validity.* (i) *With an Internal Criterion.* A factor analysis of the inter-correlations was carried out by Burt's method of simple summation. The general factor accounts for only 36 per cent. of the variance—a decidedly low figure considering the homogeneity of the test. The first bipolar factor tends to separate what Raven terms the D section of his test from the C section, and the second to separate the C section from the B. The existence of the ' difficulty factor,' found by Miss Horton, was not confirmed ; as both she and Burt pointed out at the time, the apparent presence of such a factor " may be due to the use of the tetrachoric method in correlating an unusually easy item with an unusually difficult item, since such a situation is apt to distort the assumed normality of the bivariate distribution " (cf. 6, p. 149).

(ii) *With External Criteria.* The partial correlation between the matrices test and the results for the intelligence tests, when age has been eliminated, amounts to only 0·54. This is much lower than Raven's figure, 0·855. And since tests of the analogies type are known to yield much higher correlations with children of this age, the natural inference is that some of the items have probably been badly selected or badly constructed.

Validity of the Separate Items. For the purposes of item-analysis the results obtained from the matrix test as a whole were averaged with those obtained from the other tests of intelligence (Simplex or Terman-Merrill), and the combined measurements taken as the criterion. Biserial correlations with these measurements were then calculated for every item and each age-group. The coefficients varied widely, ranging from 0·20 to 0·80, and averaging about 0·45. It was found that in the main the items in set A possess little discriminative value. With children aged 10 and 11 or thereabouts, set D appears to include the most effective items. But, with children over 12, even this set gives rather low coefficients. Items which furnished poor correlations at almost every age were A5, A8, A11, B4, B7, B8, B12, C2, C8, C10, C11, D8, D11, E3, E4, E9 : (see 10, Table XIV, pp. 74–82). It is probably the large number of comparatively worthless items in sections B and C that accounts for the unreliability of the test at the 15–30 score range, noted by Fraser-Roberts in the passage already quoted. Further calculation indicated that, if these items were omitted, the validity coefficient could be raised to nearly 0·65.

Difficulty. The proportions of children passing or failing showed that, for children over 10, there are far too many easy items and not sufficient difficult items. This perhaps is due partly to the fact that Raven's version was originally designed for use with mental

[1] We are indebted to Miss Keir for her kindness in supplying the data which were obtained under her general supervision. The present note, indeed, may be regarded as consisting very largely of an extension to her own earlier study (6).

defectives ; it would gravely impair the usefulness of the test (at least in its present form) as a means of allocatingpupils to different types of secondary school at the age of 11 plus.

The several items were scaled by the method originally suggested by Burt for grading the Binet tests (1, p. 150). This assumes that the scaled measurements may be regarded as factor-measurements for the general ' difficulty factor ' underlying all the results obtained from the different ages : the percentage passing each item at each age is first converted into a normal deviate (i.e., into a multiple of the standard deviation for the age-group) ; all the deviates for the same item are averaged, due allowance being made (where necessary) for the differing dispersions of the various age-groups. The results of the scaling revealed a well-marked bunching in the middle of the scale ; in fact, the distribution of the items approximated roughly to that of the normal curve.[1] Such a distribution is especially apt to arise when the items in a test have been chosen or devised more or less at haphazard.

Logical Construction. Investigators who attempt to draw up new tests for some particular purpose seem, as a rule, to proceed without any well-defined psychological plan. Having decided on the general form of the problem, they then merely try to think out suitable material by a rather naïve kind of free association. No one would expect to discover suitable tests by this empirical procedure in any other branch of science—say, tests for chemical substances or for changes in physiological function : every tentative test would be the outcome of a definite logical deduction.

The systematic principles available for constructing a graded or ' progressive ' test in psychological work have been fully described by Burt (9), who uses the analogies and matrix tests for his main illustrations. Accordingly, the logical structure of the several matrix-patterns included in Raven's version was carefully examined from this standpoint. Space will not permit us to report the details here ; the chief results will be found in the thesis already quoted (10, pp. 50–61). Many of the items are obviously out of place ; others are logically defective either in their internal structure or in the alternative answers supplied. Far more attention should be paid to the selection of an appropriate set of alternative answers to each item. The guiding principle should be to *maximize the information* conveyed by the child's actual response in every case. This means so constructing and selecting the alternatives that he is highly unlikely either (i) to hit on the correct answer by an ignorant guess or (ii) to choose what will be marked as a wrong answer as a result of reasoning that is, in point of fact, quite intelligent. If arbitrary marks are assigned to each item for (i) the number of relations involved, (ii) the difficulty of the relations involved, and (iii) the number of derangements, the figures so obtained for the entire series correlate with the actual order of difficulty to the extent of only 0·48 ; if, however, the items of low validity are omitted, the correlation is raised to 0·81. It thus seems clear that, in constructing the present version, the principles appropriate to the construction of such a test-series have not been applied with any accuracy or thoroughness ; and a study of the various publications describing the current form of the test suggests that it may have been rather hastily put forward (owing perhaps to the urgent demands of the war) without the usual preliminary attempt to check the value of the constituent items and to grade them in order of difficulty by appropriate statistical procedures. A better version of the test could almost certainly be constructed if additional problems were devised and tried out, and then combined with the best of the old, and the whole test systematically built up afresh in accordance with the recognized principles of test-construction.

III. SUMMARY AND CONCLUSIONS

1. Raven's 1938 version of the 'Progressive matrices' test, with other tests of intelligence, has been applied to 310 children, aged 8 to 13, and the results obtained subjected to a statistical analysis.

2. The average reliability (0·88) of the matrices test was found to be distinctly smaller than that given by Raven, and appreciably below the minimum needed for a satisfactory test of intelligence. The average validity, whether judged by internal or by external criteria, was decidedly poor (0·54).

3. An item-analysis, based on biserial correlations, showed that the validity of the separate items differed widely, averaging only 0·45 : if the poorest were eliminated, it could probably be raised to at least 0·65.

4. On scaling the tests on a linear scale, it was found that there were far too many items of medium difficulty. This seemed to be due to the fact that the test-

[1] This conclusion, obtained from a slightly larger group, confirms that obtained by Miss Keir (6, p. 146).

problems had been devised or selected in a more or less haphazard fashion. The matrix type of test-problem can be constructed on a definite logical basis, so that an evenly graded series can readily be produced, with the items steadily increasing in difficulty, according to the number and nature of the relations involved. A rigorous observance of these principles would greatly aid in improving both the validity of the separate items and the regularity of the scaling.

5. Although the test appears unquestionably promising, it would be premature to accept it in its present form, when there is so much room for obvious improvement. Instead of adopting the existing version as final, it is suggested that more items should be systematically constructed in accordance with the logical principles involved, and that an endeavour should be made to increase its reliability and validity by a systematic application of the recognized methods of item-analysis, preliminary scaling, and general standardization.

REFERENCES

1. Burt, C. (1921). *Mental and Scholastic Tests.* London : King & Sons. (Seventh Impression, 1949, Staples Press.)
2. Raven, J. C. (1938). *Perceptual Tests of Intelligence* (M.Sc. Thesis, University of London Library).
3. Raven, J. C. (1941). ' Standardization of progressive matrices (1938),' *Brit. J. Med. Psychol.*, XIX, 137–50.
4. Raven, J. C. (1948). ' The comparative assessment of intellectual ability,' *Brit. J. Psych., Gen. Sect.*, XXXIX, 12–19.
5. Banks, C. (1949). ' Factor analysis of assessments for army recruits,' *Brit. J. Psych., Stat. Sect.*, II, 76–89.
6. Keir, G. (1949). ' Progressive matrices as applied to school children,' *Brit. J. Psych., Stat. Sect.*, II, 140–50.
7. Vernon, P. E., and Parry, J. B. (1949). *Personnel Selection in the British Forces.* London : University of London Press.
8. Burt, C. (1945). ' The reliability of teachers' assessments,' *Brit. J. Educ. Psychol.*, XV, 80–92.
9. Burt, C. (1950). ' The theory of mental testing,' *Third Yearbook of Education.* London : Evans Bros. Ltd.
10. Sinha, Uma (1950). *A Study of the Reliability and Validity of the Progressive Matrices Test* (M.A. Thesis, University of London Library).

Vol. IV
Part II

The British Journal of Psychology,
Statistical Section

June,
1951

TEST-CONSTRUCTION AND THE SCALING OF ITEMS

By CYRIL BURT

University College, London

I. *A Preliminary Analysis of the Test Situation.* II. *The Measurement of Item Difficulty.* III. *Principles of Age-Assignment.* IV. *Principles of Age-Measurement. V. Alternative Procedures :* A. *Methods Based on Observed Percentages ;* B. *Methods Based on Normal Deviates.* VI. *A Worked Example.* VII. *Absolute Scaling.* VIII. *Summary and Conclusions.*

I. A PRELIMINARY ANALYSIS OF THE TEST SITUATION

The Theory of Test-Construction. Mental testing has of late become the target of vigorous attack. The objections urged some forty years ago against its introduction have been revived or re-formulated by several recent writers.[1] Such criticisms can be met in two ways. (i) In the past the advocates of mental testing have usually based their views on statistical deductions from certain psychological assumptions ; and the commonest line of attack consists in denying these theoretical assumptions, without which, it is supposed, the statistical arguments cannot be sustained. Accordingly, I propose first to point out how recent developments in statistical methodology may enable the psychologist to derive many of his procedures and his formulæ from a statement of *practical* rather than of theoretical requirements. (ii) But by far the most effective answer will be to improve the current methods of test-construction ; and that again can best be done by basing those methods on empirical rather than theoretical considerations. In this and one or two later papers I propose to describe and illustrate what I take to be the best working procedures for this purpose. The main part of the present article will be concerned with the most fundamental task of all—that of scaling a new test. But, since our critics so frequently complain that the ' mathematical psychologist ' never stops to examine the " flimsy foundations on which his procedures are implicitly based," I shall preface the general discussion by outlining, in this introductory section, what in my view is the most satisfactory approach to the whole matter.

In part, I fancy, the enthusiastic haste of many practical psychologists must be blamed for many of the strictures incurred. They are impatient of these logical niceties. To the teacher, the research student, and the medical psychologist, it seems so easy to invent a dozen or more test-questions like those described in the books, and then to mark the answers obtained in the light of common sense, and so grade the examinees for ' intelligence,' ' scholarship ability,' ' special aptitude,' or even ' neurotic tendency.' But a close examination of the tests put forward will show that, more often than not, little or no scientific attention has been paid to the proper selection of the items or to the systematic construction of the scale, and that this has greatly marred the results obtained. For instance, in four booklets of intelligence tests recently published, the order of difficulty is so erratic that I find nearly 30 per cent. of the examinees spend most of the time allotted over harder questions placed early in the scale ; in consequence they never reach some of the easier questions, which they could answer at once. The more careful investigators determine the usual coefficients for validity and (less frequently) for reliability ; but the preliminary collection of problems is still left largely to chance. In the literature reference is seldom made to this aspect [2] ; and the appro-

[1] E.g., Chambers, E. G., ' Statistics in psychology and the limitations of the test method,' *Brit. J. Psych., Gen. Sect.*, XXXIII, pp. 189–99 ; Zangwill, O. L., ' Intelligence and mental measurement,' *Introduction to Modern Psychology*, 1950, pp. 141–65.

[2] Two admirable books have recently appeared dealing with the general subject—*The Theory of Mental Tests*, by H. Gulliksen, and *Mental Testing : its Principles and Applications*, by F. L. Goodenough. Yet neither has anything to say about the theoretical principles underlying the process of testing or the compilation of test-material : (see, for instance, what topics are included and which are ignored in Appendix D of the former, on ' Sample examination in test theory ').

priate methods seem to be frequently unknown. It is urgently desirable that practising psychologists, who are now being appointed by official bodies in increasing numbers, should give closer study to the probable consequences of their procedures when applied in actual practice. And this in turn demands that the theory of the subject should be further developed with practical needs in view.

The Mechanical Theory of Communication. Owing largely to the influence of Thorndike, the process of mental testing is still commonly conceived in terms of stimulus and response. To some extent the teaching of Stout, McDougall, and the Gestalt school has led most British writers to avoid the cruder modes of description prevalent forty years ago. Yet no adequate theory or terminology appears as yet to have taken its place. And the consequence is, as anyone may observe who turns to some of the latest publications on personality testing, that the whole process is described in language which more than one reviewer has stigmatized as " vague, woolly, and metaphorical," or " lacking in all scientific precision."

Now, without assuming that the mind is nothing more than a machine, it seems perfectly legitimate for practical purposes to try to test the varying efficiency of different individual minds by the same principles that would be adopted by a modern engineer in testing the efficiency of different makes of highly intricate pieces of apparatus, say, an electrical calculating machine or a wireless receiver. But, of course, our testing-instruments would have to be far more complex.

Binet explained to the Commission appointed by the Minister of Public Instruction at Paris that his scale might be regarded as a kind of ' mental barometer ' or ' clock,' which would isolate ' general intelligence ' and measure its development on a linear scale—a time scale, in fact, since ' age ' is a temporal concept. The critics are naturally tempted to reply that we have no ground for assuming that " any psychological qualities are capable of linear measurement," and that in any case single psychological qualities " do not function in isolation." [1] Moreover, appliances like clocks and barometers are far too simple in their working to supply plausible models. I myself ventured to suggest that the operation of an effective test might rather be compared to some more complicated type of instrument—e.g., a " multiple self-registering barometer," and the mind to a computing machine that could perform problems in matrix algebra or logical deduction.[2]

The idea of an actual machine which would not merely calculate but imitate the processes of human thinking is by no means new. As long ago as 1720 the professors on the island of Laputa are said to have invented a thought-machine. In the 19th century Earl Stanhope, Alfred Smee, and Stanley Jevons all constructed working models. Jevons's [3] account is especially illuminating. His purpose, he says, was not to obtain useful solutions to concrete problems by mechanical means, but to illustrate the nature of thought-processes and the theoretical principles required to explain them. He distinguishes three main features in his apparatus : (i) the machine as such is devoid of any knowledge, but the nature and amount of the information supplied can be precisely assessed ; (ii) the machine itself is virtually " endowed with the powers of thought," i.e., it possesses the ability (within its limits) to solve the questions put to it ; (iii) it possesses the ability to communicate its solution to the enquirer in symbolic form. However, Jevons's ingenious mechanism, as Carveth Read observes, is " too perfect to represent the actual process of reasoning " ; if such a thing is to imitate human thinking, " the information supplied to it should be registered, not perfectly, but imperfectly, and the replies returned by it should be not certainly true, but only possibly or probably true." And, we may add, we shall need some statistical procedure for estimating these various probabilities.

The statistical theory of such machines has been ingeniously developed by Wiener, and applied to the interpretation of psychoneural processes. The best analogue is perhaps to be found in the working of what technologists call ' servo-mechanisms.' The function of such a mechanism is to produce an ' output-pattern ' which shall, to a specified degree of approximation, satisfy a given ' input-pattern,' in much the same sense as an explicit algebraic solution is said to ' satisfy ' a set of simultaneous equations. Now, as it happens, the development of radio-telephony and the application of electronic devices to the highly complex problems of radar have led to the development of just the kind of theoretical principles that these analogies

[1] Chambers, 1943, *loc. cit.*, pp. 190, 191.

[2] *Factors of the Mind*, 1940, pp. 90–2 ; *Psychometrika*, III, 1938, p. 160.

[3] *Principles of Science*, 1900, pp. 107f., and (for a fuller account) references, *ibid.*, p. 108, footnote 1.

require ; and more than one recent writer has suggested that such theories might well be adapted to throw light on the neurological mechanisms on which mental processes appear to depend.

If they are to represent biological systems, such mechanisms (so Wiener contends) must be based, not on classical mechanics, but on statistical mechanics.[1] Accordingly instead of energy, defined in the classical way as ' capacity for work,' he takes as his central concept that of ' information ' (a term already familiar from Fisher's use of it). Hence the efficiency of any particular machine is to be measured, not by loss of available energy, but rather by loss of available information.

In statistics ' information ' is treated, not as an intrinsic property of the informative result, but as dependent on the total relevant situation. Thus, in the theory of maximum likelihood, it will be remembered, ' information ' is expressed as the inverse of the minimum variance of any given statistic : it is in fact virtually the square of what used to be called ' precision ' ($h = 1/\sigma\sqrt{2}$) : (cf. Fisher, R. A., *The Design of Experiments*, esp. chap. XI, ' The measurement of amount of information '). In the theory of communication, the concept adopted by Wiener may be regarded as an extension of Fisher's. " Any message," he says, " may be treated as a time-series, possessing an associated quantity of information, due to the fact that the actual message is one selected from a limited number of possible messages." Thus, in the familiar Mixed Sentence test, ' A good dog defends his master ' is just one of the 720 combinations that the child can make with those six words ; nor are these various combinations equally likely. Consider first a single result or event, whose ' likelihood ' is $p = P(r|q)$: then the amount of information conveyed by that result when it occurs is $I_r = -\log_2 P(r|q)$, where I_r denotes the information supplied by any ascertained result or ' message ' r, and $P(r|q)$ denotes the likelihood of r, estimated with reference to a set of hypothetical conditions and antecedent assumptions (q), which must also indicate what is relevant. If the whole system of possible events is discrete (as in a test where the items yield duple or multiple responses), the total amount of information conveyed is expressed by $H = -\Sigma p_i \log p_i$ (21, pp. 54, 105). If the system is continuous, we may take the corresponding integral $H = -\int_{-\infty}^{\infty} p(x)\log p(x)dx$ (21, p. 54, ; 20, p. 76). Evidently such a quantity will be both positive and additive. Wiener uses logarithms to the base 2, so that the ' unit ' of information shall correspond to a probability of $\frac{1}{2}$: the unit is termed the ' bit,' i.e., ' b(inary un)it.' As thus defined, the amount of information increases as the likelihood decreases. The reader will be familiar with this principle from discussions about the scoring with the method of multiple choice : as is well known, the choice of a correct answer yields more information when the number of alternatives is large, provided the frequencies with which each is chosen in the absence of knowledge are equal : but, if the test-constructor alters the balance of probabilities in an appropriate way, the same amount of information may often be obtained with only two. For a fuller discussion of these points I may refer to ' Laboratory Notes on Test Construction ' and (23), pp. 42f. As I have there suggested, for psychological purposes, Wiener's definition could itself be extended with advantage.[2] If the different probabilities are stated explicitly, we can include not only actual facts or events, but also their presumable implications : for in mental testing we are concerned not so much with actual behaviour as with inferences from behaviour : (cf. 21, p. 97).

The theory of communicated information has been worked out most fully in connexion with telephonic communication. There what has to be conveyed is not an impulse or current whose magnitude can be specified by a single scalar number, but a structural pattern or *Gestalt* such as can be expressed more completely by a matrix of numbers. Every communication implies a *terminus a quo* and a *terminus ad quem*. Thus every test is, in a sense, a ' projection test.' Just as, in projecting a lantern-slide on a screen, we speak of a ' source ' for the projected rays, so here we may speak of a ' source of information ' : and just as for an X-ray projection we require a special receiver or screen, which ideally should impose no pattern of its own, so in testing we require an objective and impartial ' receiver.' Moreover, in all types of mechanical communication, the essential structural

[1] Wiener, N., *Cybernetics : or Control and Communication in the Animal and in the Machine* (1948). For a brief critical account of Wiener's theories, see Dr. Sholl's suggestive review in this *Journal*, II, pp. 122–4.

[2] Since this paper was written, Flanagan, in a volume just published (Lindquist, E. L., ed., *Educational Measurement*, 1951, pp. 699), has discussed in some detail ' the principal types of descriptive information derived from tests.' The very fact that he makes such constant use of the term (on some pages it occurs half a dozen times) would seem to afford independent confirmation of the importance I have attached to it in my own somewhat tentative discussions. The most urgent need, as it appears to me, is now to define the *different* ' types of information ' (on a logical as well as a statistical basis), so that the concept can be definitely helpful in arguments about the choice of experimental and statistical procedures : and here the mode of approach developed by Wiener, Wald, and their followers seems highly suggestive : (cf. also 25 and refs.).

patterns have to be transmitted by a medium. For this purpose they have first to be transformed or ' coded ' and eventually re-transformed or ' interpreted.' And not only the coding and decoding, but also the actual transmission tend to introduce cumulative obscurities into the message. The same is true in psychological communication. Accordingly, in virtue of the close analogy, the principles and the terminology developed in investigations into mechanical communication lend themselves admirably, at least as a first approximation, to the task of analysing and describing the processes involved in applying and interpreting tests. In both cases we may distinguish nine essential phases, namely (to adapt Shannon's terminology) : (i) source, (ii) original message, (iii) transmitter, (iv) signal or transmitted message, (v) interference, (vi) signal received, (vii) receiver, (viii) decoded message, and (ix) destination (21, pp. 5, 98).

But, before we can draw any inductive conclusion at all from our test-results, we must, I think assume what Russell calls the ' postulate of causal lines,' viz., " that it is possible to isolate a series of events, such that from one or at most two known members of the series an inference is possible as to the other unknown members of the series." Such ' causal lines ' are usually quite complex. And " Throughout a given causal line, some degree of constancy of *structure* must be preserved : any change in it must be gradual and intrinsic. Thus, the process from speaker to listener in broadcasting forms one such causal line." [1]

Now, if we use factorial procedures, we can often represent a structural pattern, at least to a first approximation, by two or three factors only, and sometimes even by one : hence the computational problems may be greatly simplified. In neither of the cases do the structures of the successive phases show a perfect correlation from one point of the causal line to the next, as is assumed by nearly all stimulus-response theories. As a result, there is a progressive decrease of information and a progressive increase of uncertainty, i.e., of the variance due to accidental interferences or ' chance.' This, too, we can represent by an appropriate factor. In Wiener's scheme, as will be seen from his formula, it plays the part of ' entropy.' [2] The practical task for the psychologist in constructing his tests and planning his method of assessment is thus very similar to that of the engineer in constructing (say) an efficient heat-engine : only the psychologist's problem, like that of the telephone engineer, is to prevent the degradation not of energy, but of information.

Human communication is generally reciprocal ; and, in the theory of testing, we are concerned with two distinct processes. The psychologist has to estimate the efficiency of two kinds of machinery : first, that of the test, and, secondly, that of the person or persons tested.

(i) In the latter case, which may be called the major situation, the ' source ' of the communication is the testee, and the ' recipient ' is the examiner. The examiner's object is to interpret some relevant patterned ' message ' which is generated by the source in the form of signals, and arrives more or less blurred or distorted as a result of inevitable ' interference ' with the ' causal line.'

(ii) When a test is applied, there is a preliminary communication in the opposite direction. Here the examiner must now be regarded as the ' source ' and the testee as the ' recipient.' The ' message ' must be identified, not with the test-problem itself, which is merely the carrier signal or series of signals), but rather with the solution implied in that problem as originally stated. The

[1] Russell, B., *Human Knowledge*, 1948, pp. 477, 508. Since it is the structure of the successive ' members ' that is all-important, this postulate is virtually a re-statement of Spinoza's axiom : *ordo et connexio rerum idearum idem est ac ordo et connexio idearum.* Russell's whole analysis of the processes of linguistic and logical communication is highly suggestive. Here I have ventured to combine his terminology with that adopted by Shannon and his colleagues for their work at the Bell Telephone Laboratory (21). But in some respects their statements, I fancy, need to be modified. They are concerned chiefly with audible messages which form a time-series : as experiments in television show, the *instantaneous* transmission of a simultaneous structure (e.g., that of a visible picture) involves special mechanical difficulties. Russell, indeed, suggests that it may not be possible to find a ' rule of order ' for reducing *every* ' static ' structure to a corresponding ' temporal ' structure. If so, this would greatly limit the analogy between mental and physical communication. Thus the Wiener-Shannon theory does not sufficiently allow for the fact that the mind can solve many problems synthetically and intuitively, as well as analytically and logically : (see 23, pp. 71f.). However, Shannon's more recent ' programme ' for the construction of a mechanical chess-player (26) does indicate what may be a possible scheme for overcoming the logical difficulties that arise.

[2] Wiener's use of the term ' entropy ' may be regarded as an extension of Boltzmann's well-known equation (relating entropy to probability in his attempt to re-interpret thermodynamics in terms of statistical concepts). As the choice of the letter is meant to remind us, the formula for H is closely analogous to that adopted to express entropy. And " entropy," it will be remembered, " simply means passing from a less probable to a more probable configuration " : that is, an increase of irregularity or randomness, and loss of regularity or structure ; consequently " entropy is taken to be proportional to the logarithm of the probability " (Jeans, J., *Kinetic Theory of Gases*, pp. 268f. ' Entropy and Probability ').

process of 'coding' turns the implied message into a question or 'problem'[1]: the recipient's answer forms the interpretation.

This mode of formulation seems at once more intelligible and more natural to teachers and examiners than the old attempts to describe the processes in terms of stimuli and responses or situations and behaviour. At the same time it also helps us to correct certain crude and misleading assumptions involved in their own descriptions. Too often, I fancy, both test-constructors and their critics alike forget that the primary concern of those who actually make use of mental testing is, not so much with the supposed psychological nature of the mental processes involved, as with the practical inferences to be drawn from the results actually obtained. At the same time, the general principles I have indicated enable us at once to detect and to correct the crude and oversimplified assumptions tacitly made by those who (in the words of one of them) are content to base their methods of examining and marking "on common sense and practical experience, unaided by the sophistications of the scientist." The teacher's usual mode of marking assumes that, if Tommy possesses an item of knowledge for which the test-question enquires, he will *certainly* understand the mode of enquiry, and will *certainly* supply that item of knowledge in reply, and (what is still more fallacious) that, if Tommy does not possess that item, then he will *certainly not* give the right reply. Now the foregoing analysis of the process emphasizes that it is by no means certain, but only *probable*, that Tommy will respond in this fashion. And this is but one of the many ways in which part of the information possessed both by testees and testers can be rendered unavailable through the very process of communication. Hence, if our tests are to be efficiently constructed and scored, an appropriate type of item-analysis should be made in advance to check these probabilities.[2]

Broadly speaking, the aim of the test-constructor is twofold: (i) to maximize the amount of information finally secured, while at the same time (ii) minimizing the

[1] The complete enunciation of the test-problem includes two sets of information: (i) the 'programme' to be completed (the *agendum* or the *demonstrandum*, as the mediæval logician would have said) and (ii) the 'facts' available for its completion (the *data*). In a typical servo-mechanism this information is fed in as a statement (i) of the *required* condition (e.g., the spatial position to be reached by a pilotless aeroplane) and (ii) of its *present* condition (e.g., the initial or subsequent position of the machine). The task of the machine is then to 'close the gap' (if I may borrow a phrase from the Gestalt psychology) between the required and the present conditions. As soon as that is done the machine reaches a position of stable equilibrium. When the present and required conditions are stated in terms of spatial co-ordinates, the process becomes one of steering the machine: hence Wiener's title—'cybernetics' (the art of steering or directing).

[2] A source which produces a pattern or sequence of events in accordance with certain probabilities generates a 'stochastic process.' There are two main types, according as the events are (i) independent or (ii) dependent. Evidently, if the present state of the system we are studying is dependent in part on its previous history, we are confronted with the second type of process. In physics such processes have been examined from a statistical standpoint by Boltzmann, Rayleigh, Poincaré, and others. The simplest cases occur where the present state can be treated as dependent simply on the last, and that on the last but one, and so on: such a process is known as a 'Markoff chain'; the mathematical 'expectation,' as Markoff showed, then turns out to be a weighted sum of probabilities, the weights being themselves probabilities: (*Wahrschleinlichkeitsrechnung*, 1912, pp. 299f.). In the theory of communication the sub-types that are most important are those which obey what Boltzmann (in elaborating the kinetic theory of gases) called the 'ergodic principle'; here some recurrent operation tends eventually to regularize the capricious effects of the arbitrary probabilities which we have to assume for the initial state: the most familiar instance is found in the effect of repeatedly shuffling a pack of cards.

Generally speaking, the content of any 'communication' or 'message' may usually be regarded as a 'chain' or 'causal line' in which the probability of any later portion depends partly upon the preceding. And thus the theorems relating to ergodic processes enable us to deal with many troublesome problems that arise in discussing test-responses. For example, when a child guesses the answer to a question, he is hardly ever, as current hypotheses suppose, guessing entirely at random: in the Mixed Sentences test (p. 97 above), if he chooses 'A' as the first word, it is unlikely that he will choose 'his' for the second; 'depends' is more likely, even though it does not make sense.

The important point for us is, not so much that these modes of reasoning lead to new formulæ, but rather that they show how, starting with quite *arbitrary functions* for the initial probabilities, we may ultimately reach a situation where simple calculable probabilities can be employed: at the same time they help us to specify the general conditions under which such formulæ can be validly applied. We are thus in a far better position to meet those critics who declare that psychological systems do not conform with the simple statistical assumptions required by our calculations. For an excellent discussion of the mathematical theory, see Fréchet, M., 'Méthodes des fonctions arbitraires: Théorie des évènements en chaîne' (1938), in Borel's *Traité du Calcul des Probabilités et de ses Applications*, I, III, ii (1938).

cost of securing it (I use the word ' cost ' to include time, effort, material, personal inconvenience, as well as financial expenditure). But such twofold demands are difficult to satisfy simultaneously. The theoretical enquirer, who believes in security at any price, is tempted to adopt the prudent principle of minimizing the maximum risk : and such a principle, as I have shown elsewhere, will certainly in a few restricted cases yield a reasonable solution. But in general the more practical method is to suppose the ' cost ' to be kept constant at one or more plausible levels, and then to seek that particular procedure which will maximize the amount of information obtained. On this basis the general task of test-construction and test-assessment can be formally conceived as a problem in conditional maxima. The special problems arising under this head can then be solved algebraically in the usual way by taking differential coefficients that will specify the rate at which information decreases with variations in the procedure proposed, and then making that rate a minimum. It will be observed that this is precisely the mathematical procedure on which factor analysis is based. Here then is yet another reason why, in my view, factorial methods are likely to render valuable help.

Incidentally, it will be seen, this mode of approach supplies one helpful answer to a prevalent type of criticism urged against factorial methods.[1] " We do not know " (it is said) " what it is in the individuals concerned that causes their varying degrees of success " ; consequently " the ' general factor ' usually extracted from a battery of tests must remain psychologically meaningless." Now, as I have elsewhere argued at some length, in calculating a ' general factor ' our primary object is not so much to establish the existence of some identifiable ' central ability,' or even to ascertain what is common to the variables involved, but to maximize the information extracted from what is statistically significant in the observed variations. Thus, with the psychological tests employed in the Forces, the general factor required and obtained was not identifiable with any definite mental ability, but was rather a factor giving (so far as was possible under the conditions of testing) a maximum amount of the information about each recruit's potential value as a member of the Services.[2]

Classification by Dichotomy. Factors represent structural patterns and express principles of classification ; and this is just what the theory of communicated information requires. Further, the basic factorial procedure (which seeks to maximize the information given by each successive factor) leads to a bipolar or binary type of classification. And, curiously enough, nearly all the mathematical machines based on electronic techniques proceed in this way. Thus they are really machines for classifying rather than for computing : and, when they deal with quantities, they express those quantities by means of a binary notation, that is, by a progressive dichotomous classification.

A common argument urged by those critics who (like Dr. Chambers) object to " statistics in psychology " assumes that " statistical treatment must inevitably involve quantitative measurement," whereas psychological qualities (so they insist) " cannot be expressed as measurable quantities." We now see that, in mechanical communication, measurements and numbers, when they arise, are transmitted by means of *classification*, and that classification is treated as the more fundamental way of dealing with information.[3]

[1] Cf. Chambers, *loc. cit.*, p. 195.

[2] Cf. *Factors of the Mind*, p. 19 ; Vernon, P. E., and Parry, J. B., *Personnel Selection in the British Forces* (1949). If we require information about definite *abilities*, then the selection and construction of the tests must be planned accordingly. But this will be a problem in general psychology ; in this paper I am concerned solely with problems.

[3] I gave an instance of the use of the binary method in a recent article on ' The Factorial Classification of Qualitative Data ' (this *Journal*, III, pp. 171f., and Table I). I there stated that quantitative measurements could be treated in the same way ; and a recent critic enquires how this could be done " without a considerable sacrifice of accuracy." With the ordinary Arabic notation numbers are expressed by a decimal notation. But it is well known that a number given in one scale can always be expressed in terms of any other scale, and hence in terms of a scale of 2 (the simplest possible scale). There is no loss of accuracy in such a process. Moreover, it is interesting to note that Jevons, unlike most logicians, insists that the bipolar or dichotomous mode of classification is fundamental ; and apparently was led to this view by the requirements of his ' logical machine ' (*loc. cit.*, pp. 692f.).

In individual psychology the simplest way of determining the characteristics of any individual is to proceed by what I have called ' progressive delimitation ' (*Factors of the Mind*, pp. 64, 106, 202). Thus, when I interview a pupil for purposes of child guidance and seek to assess (say) his ' general emotionality,' I approach my final verdict by a series of successive approximations, more or less as follows : " Is this child above average or below ? If above, is he also above the first quartile or below ? If below, is he also below the second octile or above ? " (The process could be continued indefinitely ; but the 8 subclasses so reached represent, I fancy, the narrowest confidence-interval of which my impressionistic judgment is capable.) When, however, I apply a cognitive test, I proceed rather as follows : " Can the child answer this easy question, representing (let us say) a mental age of 7·0 ? . . . Yes : then he is above, and not below, a probit of 3·0 (−2 S.D.), and consequently not ' mentally defective.' Can he answer a question which is decidedly harder, representing (let us say) a mental age of 7·5 (say) ? . . . Yes : then he is above, and not below, a probit of 3·2. . . . And so on." In either case, the results can be expressed very clearly in terms of a bipolar factor-matrix giving saturations for persons. With the former mode of classification (classification by what is popularly known as ' equal cuts '), the matrix would be equally balanced, and there would be one bipolar factor for each of the three questions I have put to myself ; with the latter it would be triangular (cf. this *Journal*, II, p. 112, Table V C). If (as several of my critics have objected) the substitution of classification for measurement must often result in a loss of information, let me once again insist on the double criterion : the task of the practical psychologist is to obtain (i) a maximum amount of information (ii) with a minimum expenditure of time or money.

No matter which procedure he follows, the psychologist's ultimate aim in making these assessments is to reach some practical decision. His practice therefore ought to be based on the statistics of ' decision functions ' (cf. 22).

At the child guidance centre, for instance, his duty may be to examine borderline cases of educational subnormality, and decide which are suitable for special schools and which are not. Where is he to fix the boundary ? A common rule-of-thumb is to choose the point midway between the means of the two groups. Were the numbers and the variances of two groups equal, this would at once minimize and equalize the expected number of misclassifications. Yet (cf. 5, pp. 177–8) such a procedure may prove most misleading. If in the total population the numbers of normals and of subnormals are far from equal, then this rule, when generalized, will render one type of error far more serious than the other. The statistical correction is simple : the two expressions on which the discriminating function is based have merely to be weighted in the ratio of the number in each group. Again, if it is known that the cost of educating a child in a special school is three times that of educating him in an ordinary school, then in the interest of economy a similar proportionate weighting —three to one—must be employed.[1] But in every case it is essential to ascertain and compare the probable amount of wastage which the various modes of decision will entail.

Yet, even when his immediate aim is merely a twofold classification, I readily agree that, in assessing the validity of his tests, it does not follow that a simple twofold classification will, in actual practice, provide the best way of securing the preliminary knowledge required. As I have argued in the paper just quoted, " if the psychologist's business is to find the best-weighted battery of tests to discriminate two groups on the basis of what is really a continuous variable (like ' intelligence ' or ' neuroticism '), then a discriminant function derived from a criterion that is merely qualitative (i.e., merely dichotomous) will lose much of the information that might have been preserved by the use of a graded criterion." If then we are to be ready with a convincing reply to our critics, we must also know something of the sampling distribution of our various classificatory functions.

Sequential Analysis. Throughout his preliminary investigations, therefore, the psychologist must constantly ask himself ' how much available information shall I forfeit if I adopt this experimental plan or that statistical procedure.' Every novice knows that, by increasing the number in his sample, he will increase the amount of information (defined as Fisher defines it) in direct proportion ; but too often he fails to realize that, while keeping his sample quite small, he may effect quite as large a gain

[1] Since this article was written, T. W. Anderson has published an instructive paper in which the problem of ' Classification by Multivariate Analysis ' is approached from much the same standpoint as I have adopted (*Psychometrika*, XVI, 1951, pp. 31–50). For the simplest case of all (classification into one or two equal and normally distributed populations) he reaches the same solution as I had (5, p. 177, and this *Journal*, III, p. 104). For the more complex case, in which the two types of misclassification are of different gravity, he suggests that the borderline " could be determined with sufficient accuracy by a trial and error method with the normal tables." But the proof I have given leads directly to an exact and general algebraic expression for this purpose (*loc. cit.*, p. 104). He goes on to re-examine the more general form of the problem as discussed by Rao and Slater and by Lubin (this *Journal*, II, pp. 17f., and III, pp. 90f.), and introduces several important amendments.

by refining his programme of work.[1] In the field of applied psychology the probabilities on which the assessment of information depends can as a rule be estimated only *in the course of the experiments themselves.* Consequently the practical psychologist, and more particularly the test-constructor, will be greatly assisted by adopting what is known as ' sequential analysis.'

For example, in the early days of mental testing, one of the most controversial points was the legitimacy of assuming ' special abilities.' In the traditional theory of statistical significance, there were two alternative conclusions which the investigator might draw : (i) that the results obtained are incompatible with a mere chance hypothesis : (if after all he is wrong, he may be said to commit a positive error—an ' error of the first type ') ; (ii) that the results are fully compatible with a chance hypothesis : (if after all he is wrong, he may be said to commit a negative error—an ' error of the second type '). Spearman, adopting a rather severe probability level for errors of the first type, found that the evidence supporting most of the alleged ' special abilities ' was non-significant, and so inferred that ' group factors are astonishingly rare.' Now, sequential analysis emphasizes the existence of a third possibility, namely, that *the data so far procured are insufficient for any reasonable decision to be made ;* and, to meet this situation, it proposes to introduce a ' sequential procedure of experimentation.' With this procedure we resolve at the outset to repeat our experiments until the proportionate number of cases requiring the verdict of ' non-proven ' is reduced below an agreed percentage, and to stop the research immediately that point has been reached. To do this the relevant statistical issues must be examined, not merely after all the data have been collected (as was usual with earlier researches in psychology carried out under Karl Pearson's general influence), nor yet before the experiment is planned (as Fisher rightly insists), but also during the whole course of the research.[2] With a few notable exceptions, the new type of statistical argument to which this standpoint leads does not, so far as I can see, entail great changes either in the psychologist's basic assumptions or in the formulæ that he employs. But it does supply a more convincing justification for the methods he uses. It shows how they may be refined so as to economize time and effort ; and it enables him to by-pass many of the embarrassing difficulties which the current theory of mental testing is apt to raise.

An excellent illustration arises when we consider how we can best take samples of a school population for such purposes as obtaining age-norms for tests. My own method (5, pp. 142f.) was to classify schools according to type or level—mentally defective, dull and backward, average, central, secondary, and the like ; and to take samples of each. But how many should we include in each subsample ? It is commonly supposed that the numbers in the sample should be proportional to the numbers in the total population. This is incorrect. Adopting Markoff's procedure[3] it is easy to show that a better result is obtained if we make them proportional to the product of the number and the standard deviation (thus giving more weight to the heterogeneous schools at either end of the scale). But, to estimate these proportions, a sequential procedure is necessary. Still more complicated problems occur when the tests are repeated with samples that remain largely or wholly the same (as in studying the effects of learning, test-sophistication, etc. : here (as I shall show in a later paper) the sequential problems can often be dealt with by applying the ' ergodic ' theorems mentioned above.

In applied psychology mental tests are required for two main practical purposes : first, to select a single subgroup (e.g., defectives, scholarship winners or grammar school pupils, pupils for technical schools) from the general population, usually at a crucial age ; secondly, to classify individuals of any age and of any degree of ability, with equal precision over the whole range covered. Now, with the first type of problem justice requires the assessments to be highly accurate in the neighbourhood of the borderline adopted ; and here sequential analysis shows that information is best secured, not (as is so often supposed) by a single application of some stock type of

[1] In other branches of biology, notably genetics and agriculture, where the need for economizing time and materials is imperative, the preliminary assessment of information has proved of the greatest utility. For an instructive discussion on this point the reader may refer to the final chapter in K. Mathers' *Statistical Analysis in Biology* (1943) and for suggestive illustrations to the chapter on ' Information and the planning of experiments ' in *The Measurement of Linkage in Heredity* (1938), pp. 56–68.

[2] See Wald (19). The early work of Wald and his associates was developed largely in connexion with the practical problems of war-time production ; and his book presents the method mainly as an aid to physical science. But its principles are equally applicable to psychological and educational work. In servo-mechanisms this sequential procedure can be represented by the way information as to the machine's new condition (e.g., the successive positions reached as a result of steering) is fed back from the output into the input system, though here of course the decisions are continuous.

[3] Cf. David, F. N., *Probability Theory for Statistical Methods* (1949), p. 185 and refs.

test (the Binet scale or this or that set of group-tests), but a sequence of two or more examinations, covering a diminishing range of candidates, and leading to special concentration on borderline cases only, with the aid of a wide range of sources. With this problem—that of a single binary or twofold classification—I have already dealt in previous papers.[1]

The second type of problem—apart from the occasional psychological surveys introduced by one or two authorities—has in the past figured less prominently in psychological discussions. But in the near future it will probably become increasingly important. In this latter case it is easy to show, by appealing to the principles just indicated, that the second of the two classificatory procedures mentioned above is by far the most economical ; and, by introducing certain plausible conditions, we can formally prove that *information will be greatest when both persons and test-items are graded by means of a linear scale.*

This incidentally answers the doubts so frequently raised as to whether a person's mental characteristics can be validly expressed by linear scales, and the occasional proposal to use percentile scales instead.[2] The possibility of grading tests in the same way has attracted far smaller notice. Results from tests of the more elementary processes—reaction-time, sensory discrimination, and the like—are commonly expressed in terms of ' natural units ' and a continuous scale. Tests of higher mental processes, however, are nearly all built up out of discrete items or problems, which are usually marked on a binary system (' Yes ' or ' No,' ' Pass or Fail,' ' 0 or 1 '). In this paper, therefore, I shall be concerned almost entirely with the preliminary scaling of items for tests of this type ; but, as I hope to show in a later paper, much the same problems arise, and much the same scaling methods can be used, when we are concerned with measurements and raw scores obtained from tests in their completed form. We are thus able to meet many of the difficulties that arise with the older methods of description. Edwards, for instance, in his admirable book on *Experimental Design in Psychological Research* (1950), still expresses the psychologist's problems in terms of stimulus and response (pp. 3–4) ; and among ' stimulus-variables ' he includes ' problem-situations ' and the like. But he at once goes on to state that " we have as yet no techniques for measuring such stimuli comparable to our techniques for measuring the intensity of light " (p. 3). Now, as I shall try to explain in a moment, the concept of the *difficulty* of a test is the correlate of the concept of the *ability* of a person, and thus at once provides us with the possibility of grading test-items in precisely the same way.

Here again we should, in my view, keep the practical conditions primarily in view. As Wiener emphasizes, ' messages ' of any high degree of complexity are almost bound to take the form of a time-series. If, then, we wish to construct a test-scale which will yield a maximum of information in a minimum of time, it is not difficult to prove, by adopting the principles he has developed, that the most efficient scale will be one in which (i) the items are presented to the testee in a time-order of increasing difficulty, (ii) the items vary widely in difficulty, so as to cover the whole range of ability with which the tester is concerned, and (iii) the intervals of difficulty between each item and the next are equal. It is with this threefold requirement, therefore, that the rest of this paper will be concerned.

II. THE MEASUREMENT OF ITEM DIFFICULTY

Definition. In controversies on mental testing confusion often occurs at the outset because the writers seldom explain precisely what they understand by ' difficulty.' An operational definition therefore seems badly needed. I suggest that by the difficulty

[1] *Brit. J. Psych., Gen. Sect.*, XXXIV, 1943, pp. 1–19, and *Brit. J. Educ. Psychol.*, XVII, 1947, pp. 57–71, and refs. In what follows I have in view chiefly the duties of the educational psychologist, since it is in this field that the need for such work is, at the moment, most pressing. But the same distinction holds good in other branches : for instance, the first type of problem arises chiefly in vocational selection and the latter in vocational guidance.

[2] Such doubts, for example, have been expressed by Chambers in the article already quoted. I have attempted to give a demonstration of the fact that such scales can be constructed operationally in 23, pp. 54f. For the distinction between ' graded ' and ' ungraded ' (uniform or homogeneous) tests, and their respective uses, see 5, pp. 298, 301f.

of a task we shall mean that property of a task (regarded as belonging to a stated type of tasks) which increases the probability that any individual (regarded as belonging to a stated population of individuals) will fail at it, when tested under specified conditions. This tallies with popular usage. When a teacher says " that question is too difficult for a ten-year-old boy in this school," he means that in his opinion such a boy will most probably fail to answer it correctly.[1]

Requirements. In attempting to construct a test-scale which consists of a series of tasks, we may begin by assuming, for the reasons already given, first, that the successive tasks should be arranged in order of increasing difficulty (as just defined), and, secondly, that the different degrees of difficulty should be distributed as evenly as possible over the whole range of difficulty which they represent.

The importance of this double requirement can be most readily appreciated if we consider the age-scales which are so frequently employed in child psychology. The Binet-Simon intelligence-tests perhaps furnish the best known example. In describing the early 1908 version, Binet merely says " the tests are classified according to the ages at which the majority succeed " ; but for the later 1911 revision he gives a table demonstrating the steadily decreasing proportions, and explains that in the " new order . . . we have arranged the tests according to difficulty, the easier placed before the harder." The need for making " the increase in difficulty always the same in amount " was apparently not recognized. So far as I know it was first explicitly stated in my memorandum on the English Revision (4 ; cf. 5, p. 150) ; the argument there given was that, " since the intervals between successive tests must represent equal increments of ability, they must therefore indicate equal units of measurement "). Binet goes on to explain that, in compiling his " measuring scale the general plan was to assign five tests to each mental year, and the whole was designed to run from age 3 to age 13, so as to contain 55 tests in all " ; and then, in order to assess the intelligence of any particular child, he instructs us to " take as point of departure the age at which he passes all the tests, and to add as many fifths of a year as there are additional tests passed."[2] But evidently, unless the increase of difficulty from one test to the next remains equal throughout the series, the unit of measurement (one-fifth of a year) will cease to be a constant quantity. Actual trial showed that, at any rate with London children, the double requirement was far from being adequately satisfied : in the 1911 edition, if we may judge by the number at the specified ages who succeed in passing the various items, there are as many as 12 tests which ought to be assigned to age 6 (including several placed by Binet himself at ages 4, 5, and 7), whereas there are only four tests for age 7 and only two for age 10 (5, p. 212).

Similar irregularities are observable in many other scales based on the same general plan. Now, as a rule, when test-material is first collected for a new test, it will be found that, unless special precautions have been taken, the difficulty of the items tends to conform roughly with the normal distribution. Hence there will be far too many items of the same mediocre difficulty, and far too few that are decidedly easy or decidedly hard.[3] Evidently a test-series of this form will yield a lot of superfluous information in the middle of the scale and an insufficient amount of information towards the two ends. We can in fact (with the usual type of assumptions) calculate the loss of information incurred by substituting a normal for a rectangular distribution. Thus, if the average reliability of the items (as judged by their inter-correlations) is in the neighbourhood of ·7 or ·8, the information obtained by 100 items which are normally distributed (as judged by the validity coefficient) could be got with half that number if they were evenly spaced. Consequently, what is needed in a test of the Binet or matrices type is not a normal but a rectilinear distribution. And if this is to be successfully achieved, we shall need some quantitative procedure for determining the relative difficulty of each item.

[1] Such a definition has been criticized on the ground that " the proportion of a population able to pass a particular item is no guarantee of its *intrinsic* difficulty " : for instance, " children from rural areas will find certain items relatively easier," whereas " children reared in the city will do better with others " (Goodenough, *loc. cit.*, p. 142). But the writer does not explain how ' intrinsic ' difficulty is to be measured in practice ; and she ignores the threefold stipulation that the kind of task, the kind of population, and the mode of testing must, for purposes of strict comparison, be taken into account. Arbitrary as they may appear, such stipulations are regularly included in physical science : thus the British unit of length is the distance between two points on a particular bar of bronze (not of any substance) at 62° F. (not at any temperature) and, in modern physics, even the procedure to be followed by the investigator in comparing some given object with this unit is precisely laid down.

[2] See (1) and (2) : or E. S. Kite's translation, *The Development of Intelligence in Children*, 1916, pp. 237, 282.

[3] A conspicuous case in which efficiency has been gravely impaired by neglect of this precaution is the Raven version of the progressive matrices test : (cf. Keir, G., this *Journal*, II, p. 146, and refs., especially Fig. 1).

To the best of my knowledge, the first attempt to work out a systematic method for this purpose was that described in my memorandum, cited above, on the re-standardization of the Binet tests for use in this country.[1] The method proved equally satisfactory for constructing and standardizing scales of other types, especially those in which success is scored on a pass-or-fail basis (e.g., the tests for assessing scholastic attainments, general knowledge, and special abilities, as well as the various tests for general intelligence, described in (5)). However, the principles then put forward have not escaped criticism ; and, more recently, in our own attempts to check the age-assignments of the new Terman-Merrill revision, with data obtained from London children, the whole problem has cropped up afresh.[2]

The American investigators concerned in that revision apparently think it quite impracticable to measure with any degree of precision the difficulty of the separate items. " The order of difficulty of test-items " (they say) " is not something that can be established once and for all." In their view, therefore, the utmost that can be attempted is to allocate them to their appropriate age-level. To arrange them in order *within* each age-level is held to be out of the question (15, pp. 83–4). Thurstone, on the other hand, considers that too little attention has been paid to " the logic of scale constructions." He believes it to be perfectly feasible, not only to grade test-items according to their difficulty, but also to base this grading on " an absolute method of scaling " [3] ; and has put forward two procedures for this purpose. To illustrate the underlying principle he has applied both of them to my own data for the Binet-Simon series, and has deduced a new arrangement and a new grading which, he implies, should be substituted for those published in my original report. However, as we shall find in a moment, the resulting changes are in fact extremely slight. The largest discrepancy relates to the position of test 28 (5 Numbers), which is displaced by three steps, and with his grading should follow test 31 (Distinguishing Right and Left), though it still remains within the same age-group. Thurstone's method is the only procedure [4] described in detail by Guilford (10, pp. 440–5). Hamley recommends it in his valuable brochure on *The Testing of Intelligence* (pp. 21f.). But most writers either ignore the problem, or, like Garrett (18), confine themselves to the task of assessing difficulty from a *single* group.

Fundamental Assumption. Whatever method is adopted, it is, in my view, essential to keep clearly in mind the assumptions or postulates on which it rests, and to consider how far those assumptions are fulfilled by the data to be used. Accordingly, I shall endeavour to state, as exactly as possible, the assumptions from which my own formulæ were deduced.

[1] See ' The Measurement of Intelligence by the Binet-Simon Tests,' *Eugenics Review*, VI, 1914, pp. 36–50, 140–52 ; also L.C.C. Memorandum on ' The Standardization of the Binet-Simon Scale for London Schools ' (1915), incorporated, with some abridgement, in *Mental and Scholastic Tests*, 1921, pp. 138f. and 210f. (1949 ed., pp. 150f. and 222f. ; unless otherwise indicated, my page-references will be taken from the last edition of this volume, as being now more readily accessible). Mr. Moore and I had previously applied the general method to scaling other tests and measuring differences between the sexes (*J. Exp. Ped.*, 1912, pp. 273f.). The basic idea—that of taking the median as origin and the measure of variability as the unit, and then converting observed percentages to terms of this conventional scale by using a table of the normal frequency distribution—is due to Galton (see below, p. 114).

[2] McNemar, Q. (15), pp. 87f. McNemar's criticism of my own findings relates only to a brief preliminary enquiry into the need for re-standardizing the new version for use in this country ; far more extensive data are now available, and will be briefly discussed in a joint paper shortly to be published.

[3] See (7). Thurstone's diagram on p. 449 is presumably meant to be substituted for my diagram (Fig. 21, *Mental and Scholastic Tests*, 1921, p. 139), and the order in his Table III is apparently intended as a correction of the order given in my Table III (pp. 132–3). His curve of growth, however, and the standard deviations for each year (Fig. 4, p. 443) show much the same features as mine (Fig. 22, p. 147) : but for this further problem I should now prefer to adopt the sigmoidal or ' logistic ' curve obtained with the later London Revision of the tests in *The Backward Child* (1937, Fig. 13, p. 650).

[4] To illustrate Thurstone's procedure, Guilford also takes two columns of percentages from my own results—those for ages 7 and 8 : (the figures cited are those obtained with the *original* Binet scale, as applied to London children : Dr. Copeland supposes them to be American results procured with the Stanford Revision ; but that seems due to his misunderstanding of Guilford, who does not, as a matter of fact, specify the source). In a footnote Guilford also briefly mentions the summational method (see below, p. 112, and refs.) as a " very simple device." He remarks that, " unlike Thurstone's proposal, it uses *all* the proportions." This was in point of fact the method by which my own provisional order (that indicated by the numbering in the first column of Guilford's table) was reached.

The first and basic principle follows directly from the definition I have given, and has not, so far as I know, been questioned,[1] namely, that the difficulty of each test-item must be some function of the proportion of children who succeed in passing it at a specified age (or ages). What we have next to decide therefore is the kind of function we should take. Several plausible alternatives may be suggested ; and each involves certain supplementary assumptions, differing from case to case : I shall endeavour to formulate them, as the need arises, at each step of the argument. And since the different approaches lead to different working procedures—some simple and quick, others slow and laborious, some suited to one practical aim, others to different—I shall begin with a critical survey of the various devices that have been, or might be, put forward, and briefly indicate what is the relative loss of information in each case as compared with the saving of time and labour.

III. PRINCIPLES OF AGE-ASSIGNMENT

The Methods of Binet and his Followers. One of the oldest suggestions is that of plain age-assignment. It has had a wide appeal, just because the idea sounds so simple.

Readers may recall how, some seventy years ago, the Board of Education sought to lay down ' standards ' for children in their first, second, or subsequent year at school. In formulating these standards the official ' code ' drew up a list of tasks and topics, arranged in order of increasing difficulty, and assigned them to the years for which they were deemed most appropriate. Every teacher knows what is meant by ' Standard II words ' in spelling or a ' Standard III problem ' in arithmetic ; and similarly he will describe a given pupil as being (let us say) " aged 12 last birthday, but only Standard IV in his lessons." [2] This embodies exactly the same principle as that adopted by Binet in grading his intelligence-tests and in assessing the intelligence of the individuals tested. Binet himself, however, nowhere explicitly states what this procedure implies. There appear to be at least three underlying assumptions.

(*a*) *The Reciprocity between Difficulty of Test and Capacity of Testee.* In testing individuals in the laboratory a common device is to estimate the ability of each one in terms of the hardest task that he can just accomplish with success. Thus, to measure a man's sensibility to light, we determine the faintest stimulus that he can just perceive ; and that is said to mark his ' threshold ' for visual sensation. Now this procedure implies the complementary notion, namely, that we can measure the difficulty of the task by *specifying the ability required to perform it.* In its general form the principle is by no means limited to human tasks or human capacity. Early mining engineers, for instance, found that an average horse, with rope and pulley, could drag 150 lb. of coal up a pit at the rate of 200 ft. per minute ; conversely they described such a task as a ' one-horse ' task. Indeed, as used in mechanics, the principle is implied in Newton's third law of motion ; and it seems perfectly legitimate to adopt a similar postulate [3] in the psychological field. We can, then, begin by assuming that there is a *reciprocity* between the problem of assessing the difficulty of a test and the problem of assessing the ability of a person.

(*b*) *Equality of Annual Increments.* Just as the average child, from the age of 3 to 13, adds about 2 ins. each year to his stature, so, it seemed natural to suppose, the process of mental growth would add *equal annual increments* to his intellectual capacity. From this it would follow that, up to a certain level at least, the ability of an individual could be expressed in terms of mental years. (If we are doubtful whether the rate is equally uniform between the ages of 0 and 3, and if we wish to measure the abilities of brighter persons after growth has stopped, then we can, with the Board of Education, take some other date than birth as the origin of our scale, and we may, by convention,

[1] This is conceded even by McNemar : " The per cents. passing an item at successive age-levels . . . provide the necessary data for arranging items according to difficulty, i.e., locating the items in the scale " (15, p. 82).

[2] The numbering of the standards dates from the time when children entered school at the age of 7 years. Thus to convert standards to mental ages we must add 6.

[3] Mach describes the ' law ' as a mere tautology ; Pearson regards it as a corollary deduced from more fundamental assumptions. Eddington treats it as a convention laid down for the purpose of measurement. Similar questions might be raised about the corresponding principle in psychology, and would have to be examined at length were we here concerned with the general basis of mental measurement.

prescribe some one particular year-interval—say, age 9 to 10—as defining the standard unit : these further refinements would have seemed needlessly pedantic in the early days of intelligence-testing ; but we may have to introduce them at a later stage.) Since then the ability of individual school children can be assessed in terms of mental age, it follows, by the reciprocity principle, that the difficulty of a mental test can be assessed by stating the mental age at which the typical normal child can just pass it.

(*c*) *Mental Age defined by Median Age.* The natural way to determine this mental age would be to take the average chronological age of a sufficiently large and representative group of children, chosen to form an adequate sample of the population in view. Binet seems to have started (rather as, we are told, the officials of the Board of Education did) by picking a large collection of appropriate tasks, and first arranging them roughly in batches according to their apparent level of difficulty : he then sought to determine, by trials on small groups, the average age of the children who could correctly pass a reasonable proportion of the test-items included in this or that batch.[1] Out of the experience so gained he gradually developed the idea of a systematic age-scale. At a later stage in his experiments he decided it would be far simpler to lay down a rule prescribing, not how many tests in the batch a child might be expected to pass, but how many children of a given age ought to pass each particular test. There has, however, been a wide variation in opinion as to what exact percentage should be required. Binet himself suggests that " a test may be assigned to a given age if about 65 per cent. succeed ; if 90 per cent. succeed, it is too easy " (2, tr. Kite, pp. 283f.). Subsequent investigators have adopted different percentages. Karl Pearson suggested a proportion of 55 per cent. ; Goddard, Kuhlmann, and Bobertag a proportion of 75 per cent. ; Terman at first proposed 66 per cent.,[2] but for his 1916 revision he advanced the idea of " a sliding standard, ranging from about 70 per cent. at the lower end of the scale to 60 or 55 per cent. at the upper end " ; and the same principle was adopted for the 1937 revision (15, p. 86).

The criterion that I myself put forward was the following : " To be assigned to any given year, *a test should be passed by approximately* 50 *per cent. of the children who are nominally of the age below.* Thus a test for age VI should be passed by 50 per cent. of the children aged 5 last birthday, and so on through all the series." [3] The justification for this we shall see in a moment.

At first sight the above principle might be taken to mean that we are to find the average age, and to determine it, not by calculating the arithmetical mean, but by taking the median. This has led critics to object that we are preferring the less exact measure, when some more exact measure could be found. Assuming normal distribution, the variance of the median is $\pi/2$ (= 1·57) times that of the mean ; hence the loss of information would apparently amount to more than one-third. This criticism may have some force when urged against the mode of scale construction contemplated by Binet and Thorndike, but, as will be seen in a moment, are not relevant to the method advocated here.

IV. PRINCIPLES OF AGE-MEASUREMENT

Need for More Precise Assessment. As was pointed out at the time, Binet's rather crude idea of merely allotting the items in groups to a given age is hardly precise enough for a well-constructed scale.

Consider the results that are actually obtained. Experiment shows that as many as 12 tests from the 1908 series must be assigned to age 6 : and the proportion of 5-year-old children who pass each one varies from 37 per cent. (for the test of 5 Numbers) to 63 per cent. (for that of Fingers). Thus there may be very wide differences in the difficulty of the items assigned to one and the same age. Unless we take these further variations into account, it is quite possible that at one age-level all the items may be decidedly hard (as in fact they are in age 11) and at another decidedly easy (as in

[1] Cf. (5), pp. 8f. : (as there stated, I have been specially indebted to Binet's collaborator, Dr. T. Simon, for supplementary information about the details of their procedure). Thorndike seems to have developed almost the same idea. In *The Measurement of Intelligence* (1926) he attempted to construct sets of tests for various levels of intelligence (or ' altitude of intellect,' as he preferred to call it) : the child's ability was then to be expressed in terms of the most difficult *set* of problems (of varying kinds) which he could pass, ' passing a set ' meaning that 50 per cent. of the answers were correct. (Terman and Merrill also seem to have had much the same notion, since within each age the tests are not graded in difficulty ; cf. McNemar, *loc. cit.*, p. 83.) However, it is not easy to carry this principle out in actual practice, and from a theoretical standpoint it is easy to show that the information gained from n sets each containing m problems of about equal difficulty is less than that gained from $n \times m$ problems which are uniformly graded throughout.

[2] For a brief review of these various suggestions, see 5, pp. 152f.

[3] (5), p. 152. This principle has been widely accepted, not only for revisions of the Binet scale, but for other types of test ; cf., for example, Hamley, H. R., *The Testing of Intelligence*, 1936, p. 6.

age 12). Such irregularities are bound to impair the efficiency of the scale. After all, test difficulty is a continuous, not a discontinuous variable. And by itself the mere assignment of each item to the most appropriate mental year provides no guarantee that the tests are evenly placed throughout the series, and obviously results in an appreciable loss of information and precision.

For these reasons it appeared to me essential to use fractions of a mental year, not only when measuring the ability of the child (as Binet proposed), but also when measuring the difficulty of each test. How, then, are such fractions to be computed? If, for example, test 28 (5 Numbers) is passed by 37 per cent. of the 5-year-olds and 76 per cent. of the 6-year-olds, its age-value must evidently lie somewhere between $5\frac{1}{2}$ and $6\frac{1}{2}$: but how are we to fix its exact location? It is not easy to do so without altering Binet's method of designating the age-value of the several tests.

Let us take first of all the problem of assessing the difficulty of a *single* test-item. If, in measuring the intelligence of a child, we adopt the mental year as our conventional unit, we can define an intelligence-level of (say) 6·0, 6·1, 6·2 . . . years as the intelligence possessed by an average child of 6·0, 6·1, 6·2 . . . years respectively. And, in accordance with the reciprocity principle stated above, we can use the same definition of intelligence-level when we come to assess the difficulty of a test. Now, when we are dealing with test-problems of the type envisaged here, the initial data will take the form of percentages. Hence, as already suggested, of the various methods for measuring an 'average' —i.e., of the various 'measures of location' (to use a convenient statistical phrase)—the 50th percentile or *median* will be the most appropriate to employ. We may then say that an intelligence-level of 6·0 years is a level that will be exceeded by just 50 per cent. of the age-group in question. From this we may infer that a 6·0-year test will be a test that is answered successfully by 50 per cent. of the children aged 6·0. Similarly, a 5·5-year test will be one that is answered successfully by 50 per cent. of the children aged 5·5, that is, for practical purposes, by 50 per cent. of the children aged 5 last birthday.

It is easy to show, from the theory of information, that, for any single test, the most precise way of measuring difficulty is to state the 50 per cent. level.[1] And the results thus obtained for the various tests fit in quite well with Binet's general principles. According to his injunctions, *all* the tests assigned to age 6 must be answered correctly by a given child before the child is assigned an age of 6·0; if he passes only half of them, he is awarded a mental age of $5\frac{1}{2}$. Thus the tests for what Binet calls age VI will be tests that are passed by 50 per cent. of the children aged 5·0 to 6·0. Similarly the tests for age VII will be those passed by 50 per cent. of the children aged 6·0 to 7·0. Hence the rule proposed above.

It seems clear, however, that Binet's method of labelling his tests should be amended. Consider an item of intermediate difficulty, passed (let us suppose) by half the children aged 5·5 to 6·5. To call this a test for age "6–7" or "$6\frac{1}{2}$" would be not only clumsy, but misleading. Indeed, as I pointed out in my memorandum, Binet's method of numbering the ages seems quite unnatural to the ordinary teacher; and it has certainly led to much confusion in discussing the various criteria to be adopted for deciding the age-level of each test. Instead, therefore, I suggested the following rule: "a test that will in all probability be correctly performed by just 50 per cent. of the children whose age is x is to be defined as an x-year test (whether x is a whole number or includes a fraction)." And in my own scales for testing intelligence (e.g., the Reasoning tests, Matrix tests, Performance tests, and the like, published in *Mental and Scholastic Tests* and elsewhere) as well as those for testing educational attainments, I ventured to substitute this method of age-designation [2] for that with which Binet had

[1] The proof is precisely analogous to that given by Fisher for finding the best experimental procedure for determining the toxic content of a drug: (the toxicity of the particular drug corresponds to the difficulty of the particular test: see Fisher, *Design of Experiments*, 1937, pp. 236f.). The general argument will be readily grasped if we think of each comparison as contributing one unit of information. Thus, a test which is passed by only one boy out of 100 merely discriminates that one boy from the rest, and so makes only $1 \times 99 = 99$ comparisons; a test which discriminates 50 boys who pass from 50 others who fail makes $50 \times 50 = 2{,}500$ comparisons, and it is easy to show that this is the maximum number of discriminatory comparisons that could possibly be made.

[2] Cf., for example, (5), p. 250. This method of grouping and denominating the tests is at once intelligible to the teacher, since it corresponds with that commonly adopted in so many Readers, Spelling Manuals, and lists of Arithmetical Exercises. No teacher would call the sums intended for children aged 7 last birthday '6-year-old sums.' The principle proposed in my Memorandum appears to be accepted by Guilford (10), pp. 210, 409; "a five-year-test should be one that the average 5-year-old can pass, or one that 50 per cent. of them can pass." If I seem to some readers to be labouring the

familiarized psychologists. In order to avoid confusion it was suggested that mental ages calculated according to Binet's method should be recorded in roman numerals, and those calculated according to my own should be expressed in arabic numerals (5, p. 299).

Now to re-group the children according to exact chronological age, measured in fractions of a year, and to discover the precise age-level at which 50 per cent. passed each test would scarcely be practicable. How, then, are we to determine the median age for each test? Here the analogy to which I have already referred will be instructive. In the measurement of sensory capacity very similar problems arise.[1] There, too, the ' threshold ' to be determined is regarded as a *Mittelwert* or ' average,' usually (though not always) defined as a median rather than a mean.

The older ' method of minimal changes ' tries to find by direct trial a stimulus value which coincides exactly with this median ; but in practical experiments this is now usually considered to be neither convenient nor altogether reliable. Instead, by adopting the ' method of right and wrong cases,' we determine the percentage of successes at equally spaced intervals, and from these we deduce the theoretical value indirectly by some appropriate method of calculation.[2] For example, two points placed on the skin are distinguished as two in 37 per cent. of the trials at 5 cm., in 76 per cent. at 6 cm., and so on : and we have to compute the most appropriate measurement find for the theoretical threshold. At first sight the mathematical problem that here confronts the psychophysicist would seem to be closely parallel to that which we want to solve here : and I have used the same figures above to make the analogy quite plain.

There are, however, at least two obvious differences. (i) In the simple psychophysical problem our data consist of sets of right or wrong responses given by one person to each of a number of test-stimuli, whereas in the scaling problem our data consist of sets of right or wrong responses given to one test-stimulus by each of a number of persons. Thus, the objective measurement of the stimuli (in terms of centimetres or the like) is replaced by the objective measurement of the testees' age (in terms of years and fractions of a year). In itself this is simply a change in the content of the problem not in its form ; and does not seem to entail any serious change in the working procedure.

obvious, may I remind them that the difference between Binet's method of assigning the age-values and my own has often been misunderstood in criticisms of what Dr. Shrubsall has called " Burt's 50 per cent. principle " ? And quite a number of writers appear to have inferred that the change in the test-criterion will alter the mental age attributed to the child : so that (as more than one worker in a child guidance clinic has put it) " our old I.Q.s would have to be all re-scored." From what has been said above it should now be clear that, whichever principle we adopt for labelling the various test-items, the mental age actually allotted to any child who passes such and such items and fails on others will remain the same. The question is purely one of test-nomenclature.

[1] May I emphasize this analogy in connexion with the teaching of psychophysical methods ? It has been my duty to act as external examiner for most of the psychological departments in this country : and in only one have I found students of psychology familiar with the relation between their laboratory exercises on sensory thresholds and the general theory of test-construction and the scaling of test-items. As indicated in my papers on ' Correlating Persons,' the relation can be quite simply explained. It is a special instance of the well-known ' principle of duality ' found in so many mathematical fields (cf. *Brit. J. Psych.*, XXVIII, 1937, pp. 64, 71). We start with a mark-sheet giving results for n items (stimuli, test-questions, or the like) and for N persons—say, $M = X + E$, where X denotes true scores, and E effects of irrelevant influences (chance-errors, guessing, etc.). Then, if the rows of M represent items, $R_t = MM'$ will be the product-sum matrix ' for tests ' ; and its general factor will represent the *ability* of the persons (assumed constant for all stimuli). *But M may be transposed.* Then $R_p = M'M$ will be the product-sum matrix ' for persons ' ; and the general factor will now represent the *difficulty* of the items or stimuli (assumed constant for all persons). (Students grasp the relations more readily if required to set down in parallel columns the features that correspond in the two cases, viz., (*a*) psychophysics and (*b*) psychometrics.) If now we drop the concrete designations, and simply speak of ' row-variables ' and ' column-variables ' (instead of ' tests ' and ' persons '), then one and the same set of assumptions will enable us to develop most of the mathematical theory required for *both* alternative fields. However, as indicated later in the text, the analogy must not be pressed too far.

[2] Wundt, W., *Grundzüge der Physiologischen Psychologie*, 1908, Vol. I : ' Allgemeine Grundlagen psychischer Messungen,' pp. 563f. ; cf. (13), pp. 395, 400f.

(ii) So long as we are comparing results for a single item from the test-scale with results for a single observer in the psychophysical experiment the analogy is quite close. But the calibration of a composite test, such as the Binet-Simon scale, involves a further complication. When each component item yields, not a graded measurement, but only a dichotomy (like ' right or wrong,' ' pass or fail '), it is impossible to extract a mental age from one item alone : we are bound to use a number of items, varying in difficulty. This entails certain supplementary procedures. The reciprocity principle as formulated above implies a further postulate : during the period when mental growth is approximately uniform, the most appropriate age-assignments for the component tests will be those which will ensure that, for any representative or random sample of the ordinary population, *the average mental age of the group shall coincide with its average chronological age.*

For practical purposes this forms the ultimate criterion. In theory it is not difficult to show that, under certain specifiable assumptions [1] which are usually pretty well satisfied, the general principles laid down above for determining the age-difficulty of a single test can be extended to the whole series ; and we might then hope that this criterion will automatically be fulfilled. Whether or not this is in fact the case with any particular scale is a point that will have to be empirically determined with the completed version. Nevertheless, as we shall find in a moment, the application of a number of different test-items to a number of different age-groups raises a good many further questions which must first be considered before we can accept the ordinary psychophysical procedures as applicable to our immediate problem.

However, before we take up these further questions it will be convenient to consider in turn the procedures most commonly employed, and see what their limitations are. It may after all turn out that, if for practical purposes we are content with approximate estimates, no very elaborate arithmetical devices are really necessary.

V. ALTERNATIVE PROCEDURES

A. Methods Based on Observed Percentages. 1. *Median Calculated by Linear Interpolation.* Since we have agreed to express the difficulty of a test by specifying the age at which 50 per cent. of the children tested will pass it, the simplest procedure of all will be to calculate the theoretical age-value by the ordinary method of linear interpolation as used for

[1] The nature of these assumptions will involve a consideration of the observed or assumed correlations between the component items. In this paper, however, I am concerned only with the preliminary problem, namely, the best method of measuring test difficulty of the *tests*, so as to ensure an even distribution of difficulty throughout the scale, not with methods of assigning mental ages to the *children tested.* Here it would take us too far afield to examine the validity and limitations of measuring individual ability in terms of mental years (on this, see Burt (3) and Richardson (14) and refs.). However, the disappointing results obtained with so many test-scales lead me to stress the caution given in the text, and to warn the test-constructor that, even when he has assigned an appropriate age-value to the separate test-items, he must beware of inferring, without further confirmation, that the complete test-scale will automatically yield the mental ages desired. Often it will be found that, after the test-items have been assigned as above, then (especially if the correlations between the test-items are low and variable, and no appropriate compensation is included) there is a tendency for the mental ages of the younger children to be slightly too low. The investigator will find it an instructive exercise to set out for a single year an *a priori* scheme of passes and failures such as might be expected with a set of items applied to a group of children aged 10·0 (say) whose abilities are normally distributed, showing how each child will eventually receive his correct mental age, and how each test is assigned to an appropriate fraction of a year. In practice, of course, perfect agreement with theoretical suppositions is not to be expected ; after all, the basic assumption that mental growth proceeds uniformly from year to year is itself only in very rough accordance with the facts. We are therefore bound to recognize that, whether for tests or for children, measurement in terms of mental age can at best be only a hypothetical convention ; and for the precise purposes of research some other unit, more accurately defined, will be needed. As long ago as 1915 I myself criticized the defects of the mental year as unit, and suggested a standard deviation scale (preferably derived from percentile rankings when the crude figures are not normally distributed) : cf. (3) and (5), pp. 172, 445. It is encouraging to note that, in a later article, Thurstone himself has adopted very much the same point of view (' The Mental Age Concept,' *Psych. Rev.*, XXXIII, 1926, pp. 268f.).

calculating medians.[1] If we assume that the interval between the consecutive age-groups is exactly 1·00 year, the formula may be written

$$a_i = a_l + \frac{\cdot 50 - p_l}{p_h - p_l},$$

where a_i denotes the age-value of the ith test, p_l and p_h are the proportions (for the lower and the higher of the two adjacent age-groups) which come nearest to 0·50 (50 per cent.), and a_l denotes the age at which the proportion of correct replies is p_l.

Thus for test 28 we have

$$a_{28} = 5{\cdot}5 + \frac{\cdot 50 - \cdot 37}{\cdot 76 - \cdot 37} = 5{\cdot}8\dot{3}.$$

Evidently, if exactly 50 per cent. had answered the test correctly, we should with this formula have assigned it a mental age of 5·5 exactly. The equation is thus entirely consistent with the principle laid down above.

It will be seen that to obtain an evaluation by this method, all we need and all we use are the proportions obtained from *two* age-groups only. As noted in my Report, I have found it useful alike for preliminary trials and for final checks. Professor Thurstone has since used it to obtain a fresh set of scale-values for the original Binet tests, taking as his data the percentages I myself had collected. But the values which he thus obtains differ from my own (obtained by the fuller procedure to be described in a moment) only in a very few minor instances. The largest difference relates to the test I have just taken to illustrate the method.[2] This test he finds to be misplaced by four items : according to his results it is actually harder than Test 31, which with this procedure obtains an age-value of 5·45. However, before we can decide whether Thurstone's figures are to be accepted as an improvement or not, we must first consider what degree of validity can be attached to this simple procedure.

Its defects have been repeatedly discussed in connexion with the ' method of right and wrong cases.'[3] First, it assumes that the frequency curve is a straight line, whereas in point of fact the curve is sigmoid or (to use Galton's term) ogival. In itself this is not a grave objection, since near the median the ogive will be nearly straight. Secondly, as commonly described, it leads to no measure of dispersion. This, however, could be overcome by using percentages which yield a measure of the interquartile range, though such an estimate will admittedly be rough. Thirdly, and far more serious, the formula ignores the greater part of the detailed information which, in researches of this kind, is nearly always available. In standardizing age-scales for children of school age, eight or ten age-groups will commonly be tested, and both the curve and the standardization should be determined by considering all the information supplied by all the groups.

2. *Median Obtained by Graphical Interpolation.* To some extent the difficulties involved can be met by substituting a graphical and curvilinear interpolation, based on all the percentages, instead of an arithmetical and linear interpolation based on only two.

The chronological age of the successive groups is marked off on the base line, and the percentages are plotted as ordinates. The points can then be fitted with a smoothed sigmoid curve, and the median found in terms of age. The curve so drawn need not be a normal ogive ; and thus allowance can be made for those cases in which the standard deviations of the age-groups differ. As Yule points out, the graphical method can be highly accurate, " provided we can assume careful and accurate draughtsmanship " : (he advises the use of a flexible spline).[4] This procedure is almost as simple as the first, and so has much to recommend it for preliminary purposes. The graphs themselves are always instructive, and provide useful checks, even when the age-values are calculated by more rigorous arithmetical procedures.

[1] Cf. Yule (11, 1910 ed.), p. 117. In psychophysical work the use of this formula for determining a sensory threshold was apparently first introduced by Wundt and his students in 1886 (cf. *Physiol. Psych.*, I, 1908, p. 594 : see also Guilford (10), pp. 41, 168, and Woodworth (13), p. 407). In the formula above I have adopted Guilford's notation (*loc. cit.*, eq. 72) slightly modified.

[2] *Loc. cit.*, Table V, p. 451.

[3] Cf. (6), p. 60, (10), pp. 168f., and (13), pp. 408f.

[4] Yule (11, 1910 ed.), p. 118 : cf. also Woodworth (13, p. 405 : ' Graphic treatment of data '). When the standard deviations show no regular variation, the quickest and most instructive graphical procedure is to plot the figures on ' normal probability integral graph paper,' and then draw the best fitting straight line. Since so many of the methods in actual use are based on the assumption of linear relations, such graphs provide a quick and simple way of testing whether this assumption is justified by the data.

3. *Mean Calculated by Summing Percentages.* Nevertheless, though graphical methods save time and trouble, they are unsuitable for rigorous enquiries. If we desire a simple arithmetical procedure which will take into account all the available data, the most obvious suggestion would be to calculate, not the median, but the ordinary average (i.e., the arithmetic mean), and assume that the mean and the median will coincide as closely as is required for practical needs. By subtracting successive percentages, we could, if we wished, obtain a frequency distribution of the familiar kind, and then compute the mean in the ordinary way. But in point of fact actual subtraction is unnecessary : precisely the same result can be obtained by summing the observed percentages, and deducting the total from the borderline-age that terminates the series.[1]

The method is shown in Table I (columns 2 and 4), where the data are taken from my original memorandum (5, Table III, pp. 144–5). For test 28 the total percentage is 717·3, or, in terms of the proportions, 7·173. This figure is then subtracted from the age marking the end of the series, namely, 13·0 (since age 12·5 ranges from 12·0 to 13·0). We thus obtain, as the age-value of the test, 13·00 — 7·17 = 5·83. Similarly, for test 31 we obtain an age-value of 5·90. By this procedure, therefore, and contrary to Thurstone's finding, test 31 proves more difficult than test 28. And the same method can be readily applied to calculate the standard deviation : (for a worked example, see 5, p. 442).

TABLE I. DETERMINING DIFFICULTY IN YEARS (*Yr*) AND PROBITS (*P*)

Age	Test 28. 5 Numbers		Test 31. Distinguishing Right and Left	
	Percentage	Standard Measure	Percentage	Standard Measure
3·5	1·2	−2·26	0·6	−2·51
4·5	15·2	−1·03	11·4	−1·21
5·5	37·3	− ·32	52·0	+ ·05
6·5	75·7	+ ·70	75·7	+ ·70
7·5	94·0	+1·55	83·8	+ ·99
8·5	95·6	+1·71	90·1	+1·29
9·5	98·8	+2·26	97·1	+1·90
10·5	99·5	+2·58	99·8	+2·88
11·5	100·0	—	100·0	—
12·5	100·0	—	100·0	—
Total	717·3	5·19	710·5	4·09
Average	—	·65	—	·51
End of Series	13·00	5·00	13·00	5·00
	7·17	·65	7·10	·51
Difference	5·83 *Yr*	4·35 *P*	5·90 *Yr*	4·49 *P*

Now, since for every test the total proportion is subtracted from the same terminal figure (13·00), these total proportions (or the average percentages derived from them) will furnish exactly the same order, and exactly the same relative distances, as the mean mental ages. Hence if we merely require an appropriately spaced order, the final step is unnecessary ; and there may be advantages in omitting it. To teachers, unfamiliar with the idea of ' thresholds,' the attempt to express the difficulty of a test by giving it a mean or median mental age often proves a little puzzling, and the confusion is

[1] The justification for the method is explained on p. 442 of *Mental and Scholastic Tests.* It is a little puzzling to find Guilford stating that with this method " the probable errors are determined with great difficulty " (10, p. 209) ; the cumulative method can be adopted quite as easily for determining the standard deviation as for determining the mean. The method is due to Sir G. Hardy, and is fully described in Palin Elderton's *Frequency Curves and Correlation,* 1906, pp. 19f. (1938 ed., p. 21). In German laboratories the method has frequently been used for determining sensory thresholds ; and is there known as the *Vollreihenmethode* (" complete series " : cf. Abderhalden, G., *Handbuch der biologischen Arbeitsmethoden,* 1927, Abteil. VI, Teil A).

increased when the ages so calculated do not tally with those allotted by Binet himself. Accordingly, in such cases it may prove simpler to explain that the items have been arranged in the order of the *average percentage passing*. This was the course adopted in rearranging the tests of the original Binet scale in an order of difficulty suitable for London pupils and in comparing the different orders obtained from different groups of children—normal, mentally defective, and the like.[1]

Nevertheless, convenient as this procedure is for preliminary scaling, it has definite defects as a device for exact research. First, as its German name implies, the *Vollreihenmethode* assumes that the 'series' is 'complete.' But, in scaling tests, as in calculating sensory thresholds, the tail percentages are frequently missing; and we are thus confronted with what is commonly known as the 'cut tail error.' In my own investigation of the Binet scale, the age-groups ranged from 3 to 14; and therefore, except for the easiest and hardest tests, the data were fairly complete. But in most enquiries a far smaller number of groups is tested, and the percentages available are usually curtailed, like the rows in Table II below. To meet this difficulty the so-called 'double-step device'[2] may occasionally be used. But in my experience it is apt to produce untrustworthy results; and some more rigorous procedure seems to be badly needed.

Secondly, in dealing with children, the standard deviations of the individuals from successive age-groups increase from year to year. Consequently, when plotted against chronological ages, the row of percentages for any given test does not form a normal ogive. As will be seen from Table I, the percentages below the 50 per cent. are apt to be compressed, and those above that level unduly drawn out. The frequency curve is thus markedly skewed; and consequently the mean does not coincide with the median.

B. Methods Based on Normal Deviates. 4. *Median of Best Fitting Ogive: (The Constant or Probit Method).* Difficulties like the foregoing have long been recognized in experiments on the determination of sensory thresholds. In that field of work it is now generally admitted that the most rigorous procedure consists in fitting a theoretical sigmoid curve to the observed frequencies by an indirect use of the principle of least squares, and calculating the mean (or median), and (when desired) the standard deviation, by the appropriate formulæ.

As we have seen, in scaling a test-item the successive chronological ages of the children tested can be regarded as forming a fixed or constant series, similar to that presented by the successive values of the stimulus in determining a threshold for skin or weight discrimination. Hence we can, with very little modification, adapt the procedure worked out by G. E. Müller for what he termed the 'Constant Method.' The steps will be much the same as those described in books or chapters on 'psychophysical methods' (6, pp. 68f., 10, pp. 176f., 13, pp. 408f.).

Since percentages and percentiles cannot be treated as a basis for measurement on an additive or linear scale, the first step is to convert them into multiples of the standard deviation by means of the normal probability function.[3] (To avoid negative deviations, 5 is sometimes added to the standard figures; they are then said to be expressed in 'probits,' i.e., probability units: cf. 17.) This assumes that the underlying characteristic measured by the tests is distributed, approximately at least, in accordance with the curve of normal probability.[4] The conversion is illustrated with tests 28 and 31 in Table I, columns 3 and 5.

[1] Cf. *Mental and Scholastic Tests*, pp. 145, 147, *et seq.*

[2] Cf. Woodworth, *loc. cit.*, p. 405.

[3] Since (in the usual notation) we require to find the appropriate x for given values of p and q, not vice versa, the most convenient form of the table is that given in Appendix C to Kelley's *Statistical Method*. Mr. Chambers complains that so many of our procedures wrongly assume that test-results must always follow a normal distribution. But I have myself pointed out that, with the older and younger age-groups, the Binet tests are apt to yield skewed distributions, and have indicated how such situations may be dealt with. They can be transformed to an approximately normal distribution, or better still (if extreme accuracy is desired) be fitted by some curve of a more general type. As is well known, the normal distribution is derived from the discontinuous but symmetrical binomial. The more general form embraces the asymmetrical binomial; but even in this the contributory causes are independent. The most general form of all, therefore—at least for the psychologist's purposes—is that based on the hypergeometric series or the corresponding continuous curves (Pearson's Type I or in some circumstances Type III): to calculate the deviates from the percentages is not so laborious as might be assumed; for small groups Greenwood's Tables, and, for large, Pearson's table of the beta-function, may be used: (cf. *The Backward Child*, pp. 665f.).

[4] In psychology the earliest use of "Gauss's function" (as he terms it) is due to Fechner, who introduced his famous 'Fundamentaltabelle' in *Die Elemente der Psychophysik* (1859; 1888 ed., p. 108). The adoption of the least squares procedure for fitting the ogive, and determining the median and standard deviation, is due to G. E. Müller (*Die Gesichtspunkte und die Tatsachen der psychophysischen Methodik*, 1904 (esp. para. 11: for the weights, see *id.*, *Pflüger's Arch.*, XIX, 1879, pp. 204f.).

The earliest use of normal deviates to measure the characteristics of different *individuals* is due to

The simplest method of getting a single value for each would be to average these deviates just as they stand : (see foot of columns ; we note that the result agrees with the preceding method in making test 28 the easier of the two). That does not give a difficulty-value in terms of a mental age. However, the detailed figures provide us with a number of empirical equations (eight per test in Table I) for connecting the normal deviates to the corresponding age-values, viz., $3{\cdot}5 = M - 2{\cdot}26\ \sigma$ and so on. These equations can be solved by the method of least squares (or some equivalent procedure) so as to determine the best fitting values for M (the mean or median) and σ (the standard deviation). With this procedure the more extreme percentages, such as 0·6 and 99·8, will evidently exert considerable influence on the final figure, since they yield values for the deviates that are disproportionately large. For the same reason a very small change in such percentages will produce a very great change in the corresponding deviates. We must therefore regard these extreme deviates as the least reliable ; and to avoid any undue influence we must compensate for the effects of the transformation by weighting each deviate differentially according to the information it supplies, i.e., according to what might loosely be called its reliability or precision.[1] That can be effected by using the system of weighting introduced by Müller and modified by Urban. Müller's weights are simply z^2 (where z is the ordinate corresponding to the proportions p and $q = p - 1$) ; Urban's are $\frac{z^2}{pq}$, as may easily be verified by comparing them with the figures in the Kelley-Wood table. In my opinion, however, it would be better still to follow the mode of calculation described by Fisher (16, pp. 8–11, 17). The complete probit method is deduced from the principle of maximum likelihood instead of from that of least squares, and should thus yield a maximum amount of information. It has the further merit of providing a more rigorous estimate of the statistical significance of the values obtained and of yielding (if required) their ' confidence limits.' Fisher's weights, it will be seen, are identical with the Urban-Müller weights (except that the latter are usually given in the form of *proportional* weights), and are deduced from estimated ' working probits,' instead of from the observed values. The actual computations I need not illustrate, since I have given a fully worked example elsewhere.[2]

If then the greatest possible information is required, this procedure is probably the best, so long as it is legitimate to assume that the standard deviations for the various groups are approximately equal. In the past this further assumption has been freely made.[3] But, when we are dealing with the younger age-groups, it would seem to be scarcely valid ; and appears doubly questionable when we are scaling tests for general intelligence. The effect of the differing variabilities is that the sigmoid curve obtained by smoothing the percentages is not a normal ogive. Consequently, the best fitting line, expressing the regression of the test on age, is not a straight line, but, as a rule, appreciably curved.

Galton (*Hereditary Genius*, 1869), who prints an analogous table (taking the requisite figures from Quetelet). In the scale there set out, the unit is $\frac{1}{\sqrt{2}}\ \sigma$, i.e., it is based on the ' modulus of precision ' ; but later Galton kept to the quartile deviation or probable error. This was adopted from Galton by Cattell, and so for long remained the preferred unit with American writers, e.g., Thorndike and his followers. It is still the basis of the method of test-scaling described and illustrated by Garrett (18, 1949, pp. 147f.). Galton's essential principle is formulated by Kelley even more clearly than by Galton himself : " Galton," he says, " considered two measures, which attempt to measure the same function, to be comparable, (*a*) when each was expressed as a deviation from the median of the group from which the person measured belonged, and (*b*) when each such deviation was divided by the quartile deviation of that group " (*Statistical Method*, 1923, p. 114).

The change from the quartile deviation to the standard deviation was due to Pearson. So far as I know, the first proposal to employ it for assessing intelligence, as tested by the Binet pass-or-fail tests, was put forward in my paper of 1914 (3).

[1] Here again a proof is given by Fisher for the familiar case in which the probit method is used to determine the toxicity of a drug. As he puts it, " the quantity $I = nz^2/pq$ is used as the weight to be assigned to the estimated value of x, when a number of tests at different dosages are to be combined " : in this equation I denotes ' amount of information ' and n the number of individuals tested (*The Design of Experiments*, p. 236).

[2] (5), p. 448. At my suggestion Miss F. E. Sharp tried this procedure with the Terman revision. On the whole, the age-values differed very little from those obtained by the shorter method described above.

[3] It seems tacitly implied in the sections on ' The Scaling of Test-Items ' in such text-books as Garrett's (18 ; cf. also McCall, W. A., *How to Measure in Education*, 1929, chap. X, pp. 272f.). It should be noted, however, that American investigators who use these procedures with pupils of different ages generally group the children, not by age, but by grade ; and (in many schools at least, as the data reported indicate) the standard deviations for successive grades do not change so markedly as the standard deviations for successive age-groups.

5. *Test Difficulty as a General Factor.* As I have already indicated, most of the methods we have been considering up to now are suited primarily for cases in which there is only a single series of cumulative percentages obtained from just one test-item. But in a composite test, like the Binet scale, there may be as many as 50 or 60 items ; and we may have, from each age-group, almost as many percentages as there are test-items. Now all the items have been selected because they belong to the same general class (e.g., tests of intelligence as Binet conceives it) : they must therefore all be saturated by the same general factor. In such a case it follows from our definition of difficulty that we should attempt to assess the difficulty of each constituent item, not by its regression on some external criterion (the chronological ages of the children tested), but rather by the internal evidence procured from the test as a whole. And for this purpose we must adopt the standpoint taken when we ' correlate persons,' not that taken when we ' correlate tests.' In other words, what we have to compare or covariate are not the variables expressed by the *rows* of tables like Table II, but the variables expressed by the *columns:* cf. p. 109 above, footnote 1).

When we correlate tests, the factor-measurements are measurements describing the individual *persons ;* when we correlate persons (or groups of persons) the factor-measurements are measurements describing the individual *tests.* And with this procedure, as I have endeavoured to show elsewhere, the general factor will be a factor that represents the *difficulty* of the several tests, just as in the more usual approach it represents the *ability* of the persons tested.[1] Hence, if we are guided by the principle of maximizing information, *the ideal method for assessing the difficulty of each test-item will be to calculate its ' general factor-measurement,'* that is, the measurements for that particular factor which is common to all the persons (or to all the groups of persons) correlated, and which accounts for the *greatest proportion of the variance.* Accordingly, having transformed each column of percentages into a corresponding column of normal deviates, our task will be to seek the ' line of closest fit ' for these deviates. Since it is the individuals in the separate age-groups that are assumed to be normally distributed, it will be the *columns* of percentages that form a normal ogive, and therefore it is the columns of deviates which we may more plausibly expect to mark points on a linear scale.

Let us then start once again by considering first of all a single age-group, a say (where a may denote any one age-group out of n). Let c_a be the degree of difficulty possessed by a hypothetical test which is passed by exactly 50 per cent. of the ath group, the difficulty being expressed in mental years. Then, since the distribution is normal, c_a will be the average mental age of that group ; and, since its average mental age is to coincide with its average chronological age, c_a will also be the group's average chronological age. Further, let σ_a denote the standard deviation of the group (i.e., the standard deviation of the mental ages of the individual members) measured in the same units as c_a. Now let t be an actual test in which a proportion of p_{ta} succeed (out of the entire number in the ath age-group), and a proportion of $1 - p_{ta} = q_{ta}$ fail ; let d_{ta} be the normal deviate corresponding to a proportion p_{ta} ; and m_{ta} be the difficulty of test t as inferred from the proportion of children in the ath group who pass that test (p_{ta}), and from the known or deducible facts about their mean mental age (c_a) and dispersion (σ_a). In accordance with our reciprocity principle, m_{ta} (the difficulty of the test) will be measured in the same units as c_a and σ_a (in other, in the same units as we are adopting to measure

[1] In criticizing my original proposal to calculate factor-measurements from ' correlations between persons,' in order to obtain ' difficulty ' as well as ' content ' factors, Dr. Stephenson and others have raised one or two interesting objections. The first and the commonest is that " factors represent *abilities* of persons and ' difficulty ' is not an ability " ; therefore, we are told, it must be an artefact. It is perhaps sufficient to say, in reply, that few statisticians would limit the concept of factor in this way. Secondly, it is said, " here, as in correlating tests, we should use reduced communalities." I would reply that their use depends on the question to be answered : for certain questions, I fully agree, we should allow for specific factors. Thirdly, we are told, " since tests are not normally distributed, the familiar weighting equations are no longer valid ; on the other hand, as Thomson (p. 53) explains, the introduction of specifics prevents us from finding exact factor-measurements such as would yield an exact reproduction of the scores. Hence there is no method left which can claim any plausibility." With this I do not agree at all. (*a*) In spite of the common impression to the contrary, it is quite possible, even when specific factors are introduced, to find " exact factor-measurements " to fit the usual requirements. In some ways it is easier : for, when correlating tests, we assume that persons who have the same test-measurements should have the same factor-measurements, and of course psychologically this is, to say the least, somewhat questionable : different persons reach the same answers in different ways. But tests do not possess this flexibility. Hence it is quite easy to introduce supplementary conditions which will render the solution unique. (*b*) It is, I admit, widely supposed that " in correlational work a least squares solution is not valid, unless the variables are normally distributed." But, by using Markoff's theorem on least squares, it is quite possible to derive the least squares solution without any assumption of normality : (cf. Neyman, J., and David, F. N., ' Extension of the Markoff Theorem,' *Statist. Research Mem.*, II, 1938, pp. 105f.).

the mental ability of the children), and from the same origin : for example, the units can be years, and the origin can be the child's age at birth. Then, it follows from our previous assumptions that

$$m_{ta} = c_a + \sigma_a \cdot d_{ta} . \qquad \text{(i)}$$

If we take some other age-group, a' say, we shall similarly have

$$m_{ta'} = c_{a'} + \sigma_{a'} \cdot d_{ta'} . \qquad \text{(i } bis\text{)}$$

Now, if there were a general factor of difficulty common to all ages, and if (as would be necessary for a perfect test-instrument) there were *no other* influences affecting the success or failure of the children, then for every age and test we should have $m_{ta} = m_{ta'} = m_t$ (say), and therefore

$$d_{ta} = \frac{m_t - c_a}{\sigma_a} ; \qquad \text{(ii)}$$

and this gives for the standard deviation of d_{ta} (within a single age-group)

$$s_a^2 = \frac{1}{N} \{ \Sigma d_{ta}^2 - N \bar{d}_{ta}^2 \} = \frac{1}{\sigma_a^2} s_m^2 ,$$

where s_a denotes the standard deviation of d_{ta} (i.e., of the deviates) and s_m denotes the standard deviation of m_t. From this we have at once

$$s_1^2 \sigma_1^2 = s_2^2 \sigma_2^2 = \ldots = s_n^2 \sigma_n^2 \text{ for all ages.} \qquad \text{(iii)}$$

Actually,[1] however, since there always will be other influences, the values obtained for m_{ta} from the different age-groups will not agree precisely, and each value will differ somewhat from the ' true ' value (if there is one) : so that we can, if we wish, put

$$e_a = m_{ta} - m_t, \qquad (a = 1, 2, \ldots n),$$

where e_a denotes the divergence between this ' true ' value (m_t say) and the value based on the result from the ath age-group.

Our problem, therefore, may be stated as follows : given n different values for the difficulty of each of our N tests, how are we to combine the n values so as to minimize these divergences, and thus arrive at the best possible estimate in every case ? Let us assume that we can make a reasonably satisfactory estimate, $\bar{m}_t$ say, by using some simple linear function.[2] We may then imagine the n sets of values that we shall obtain for each of the N variables, $m_{1a}, m_{2a}, \ldots m_{Na}$ $(a = 1, 2, \ldots n)$, to be represented by a system of points ; and our task will simply be to find the line of closest fit to this system. Accordingly, let us begin by writing

$$\begin{aligned} \bar{m}_t &= b_1 m_{t1} + b_2 m_{t2} + \ldots + b_n m_{tn} + A \qquad \text{(iv)} \\ &= b_1(c_1 + \sigma_1 d_{t1}) + b_2(c_2 + \sigma_2 d_{t2}) + \ldots + b_n(c_n + \sigma_n d_{tn}) + A \\ &= b_1' d_{t1} + b_2' d_{t2} + \ldots + b_n' d_{tn} + A' \\ &= b_1'' \frac{d_{t1}}{s_1} + b_2'' \frac{d_{t2}}{s_2} + \ldots + b_n'' \frac{d_{tn}}{s_n} + A', \qquad \text{(v)} \end{aligned}$$

where $b_a' = b_a \sigma_a$, $\quad b_a'' = b_a \sigma_a s_a$, $\quad$ and $A' = A + \Sigma b_a c_a$. $\qquad (a = 1, 2, \ldots n)$

Since there are N tests we shall have N such equations to determine the n best weights, $b_1, b_2, \ldots b_n$. Thus stated, the problem becomes that of finding general factor-measurements by the method of weighted summation, with $b_1, b_2, \ldots b_n$ as the regressions of the several age-groups on the component common to every age. So far, therefore, the procedure required would seem to be simply Pearson's

[1] It is at this point that Professor Thurstone appears to part company. He in effect assumes that $e_a = 0$ *always*, i.e., not only for every age-group, but for every conceivable investigation. This follows from his statement that the correlation between the deviates for the N tests as obtained from any two age-groups (i.e., between d_{ta} and $d_{ta'}$, $t = 1, 2, \ldots N$) can always be taken as 1·00. It is true that in the particular investigation by which he illustrates, the correlations do not fall greatly short of 1·00. But it seems hardly legitimate to assume *at the very outset of the argument* that the errors (e_a) can always be ignored.

[2] In assuming that the distributions of the results obtained from two (or more) groups can be represented as normal distributions on the same base-line, we are in effect assuming a linear relation : for it then follows that, when the deviates of one group are plotted against the deviates of another, the points should lie approximately on a straight line. With well-constructed tests, the high correlations usually obtained between orders of difficulty for different groups provide a sufficient guarantee of *approximate* linearity. When the correlations are not so high as usual, the possibility of curvilinear relations can conveniently be investigated by plotting the deviates from *all* the groups against their averages or sums. And if (as we hope) the general factor of difficulty is the most important factor determining the figures observed, then the lines so plotted should be very nearly ' synclinal,' i.e., they should converge towards a common point on or near the base-line.

method of finding the ' line of closest fit ' by the aid of the correlations or covariances between the ages with complete variances inserted in the leading diagonal of the covariance matrix.

There is, however, at least one minor complication. The deviates d_{t1}, d_{t2}, etc., are not measurements directly observed ; and, as we have seen (p. 114), they have varying degrees of precision. In such cases, it is generally held, we should, for any rigorous estimation, first ' prepare ' our initial equations by an appropriate form of weighting, before we embark on the usual routine calculation ; the principle has often been stated : " observations of different weight can be treated as observations of unit weight if each equation of condition is first multiplied by the square root of its weight." [1]

How then are these weights to be computed ? With the data that confront us here the problem is somewhat different from those usually dealt with in the theory of errors. We can, however, fall back on our fundamental requirement that information should be maximized. It is then easy to show, by the usual type of argument, that the most efficient [2] estimates will be obtained if the weight of each equation is taken to be inversely proportional to the variance of the residuals, V_t (say) ; and V_t in turn will be the sum of the variances of the component terms, namely, of

$$v_{ta} = \frac{p_{ta}\, q_{ta}}{z_{ta}^2} \cdot \frac{\sigma_a^2}{N_a}, \qquad (a = 1, 2, \ldots, n) \quad \text{(vi)}$$

where p_{ta} (as before) is the proportion of the ath group passing test t, $q_{ta} = 1 - p_{ta}$, z_{ta} is the corresponding ordinate, N_a the number in the ath group, and σ_a its standard deviation. If there were only one age-group, the weights would be $1/v_{ta} = \frac{z_{ta}^2}{p_{ta}\, q_{ta}}$, $(t = 1, 2, \ldots, N)$, a form of weighting used (as we have seen) both by Urban and by Fisher.

This procedure seems to me by far the best in theory. But the calculations will be exceedingly arduous. It remains therefore to see whether, by invoking certain simplifying assumptions (such as would be approximately fulfilled in the commoner type of case) we cannot reduce it to a more practicable form.

6. *Test-Difficulty Determined by Simple Summation.* In a previous paper I have shown that, provided (i) any irrelevant or unintended weighting has been excluded and (ii) the best theoretical weights do not differ greatly in size, comparatively little change is made in the final figures if *equal* weights are substituted for differential weights, i.e., if the decision to introduce weighting is abandoned altogether.[3] And here, as elsewhere, it is not difficult to show that very little information is likely to be lost if we omit the least reliable deviates and substitute simple summation for weighted summation.

Let us consider first the supplementary weights introduced to allow for the varying precision of the deviates $(1/V)$. If (as is the general practice in this country) the tests are standardized by being applied to *complete age-groups*, the number of children in the several groups, N_a, will be very much the same. This figure can therefore be omitted from equation (vi) without any appreciable loss. Further, except when p_{ta} or q_{ta} are small (say, less than 10 per cent.) the values of $\frac{p_{ta}\, q_{ta}}{z_{ta}^2}$ do not vary greatly. If, therefore, we omit the few deviates that are based on the more extreme percentages, we can drop these ratios also. We shall thus reduce V to $\sigma_1^2 + \sigma_2^2 + \ldots + \sigma_n^2$: and this sum will be the same for all the equations.

Now, with any well-constructed battery of test-items the correlations between the various age-groups will be pretty high. With the Binet tests, for example, they range from 0·95 to 0·98. In such a case the differences between the regressions $b_1, b_2, \ldots b_n$ will be exceedingly small, and $s_1\sigma_1$, $s_2\sigma_2, \ldots,$ and $s_n\sigma_n$ will all be approximately equal. Under these conditions and with these successive simplifications, we can now write,

$$\bar{m}_t = B(d_{t1} + d_{t2} + \ldots + d_{tn}) + A'. \qquad \text{(vii)}$$

Here B and A' will be the same for every test, and can be so chosen that $\bar{m}_t$ is measured from the origin desired and in the units desired. Or indeed we can omit these two scalar constants and estimate the relative difficulty of each test by simply summing the normal deviates. *All we have to do is to*

[1] Whittaker, E. T., and Robinson, G., *The Calculus of Observations*, 1924, chapter on ' The Method of Least Squares,' p. 222.

[2] The term ' efficient ' is here used in Fisher's sense : ' efficient statistics ' are those that (with large samples) " make use of the whole of the relevant information which is available " (Fisher, *loc. cit.*, p. 285).

[3] ' The Influence of Differential Weighting,' this *Journal*, III, pp. 111f.

convert the observed percentages to normal deviates and take the sum or average of these deviates as they stand. This was, in fact, the method proposed in my original memorandum.[1]

7. *Test Difficulty expressed in Mental Years.* The numerical expressions thus obtained for the difficulty of the several tests are in terms of a somewhat vague and arbitrary unit. For research purposes that is sufficient. But, as we saw at the very outset, the practical psychologist is concerned, not merely with the collection of knowledge, but with its communication ; and if his function is to report to others (who are not specialists like himself) the information he has obtained about his tests and test-results, that information should be conveyed in terms of some concrete and intelligible unit—a ' derived ' unit, therefore, rather than a ' raw ' unit. Where the tests are applied to children, by far the most useful of such units is the ' mental year.' It is very largely the use of this simple and plausible interpretation that has rendered the Binet scale so popular.

The attempt to measure tests in terms of mental ages has received even more criticism than the attempt to measure individuals in those terms ; but it can be defended on several grounds.

(i) The requirements of communication demand that our numerical statements should convey as much information as possible. This means that, in order to facilitate the interpretation of our test-results, we should give each test as much direct meaning as we legitimately can. And, for the purpose of describing degrees of difficulty, no concept is so readily and so widely understood as that of mental age.

(ii) The transformation from raw units to units which are equivalent from one test to another permits us to compare the difficulty of items belonging to different scales, or to tests for different subjects and different abilities.

(iii) Owing to the common system of annual promotion (as well as for other reasons) it has been found convenient, from the time of the old Board of Education code, to express norms or standards of educational attainment in terms of mental years ; and even tests of mental ability and development (like those of the Binet scale) often consist of items grouped according to age-levels. In such cases the grouping of items and classification by ' standards ' can be effected far more easily and accurately if the difficulty of each item has first been measured in terms of a continuous (if conventional) age-scale.

(iv) More generally, since there is a reciprocal relation between the abilities of individuals and the difficulty of the tasks they can perform, it is helpful to retain the same unit for both.

(v) Recent research has shown that, when measurements of relative difficulty are expressed in age-units by means of an adequate scaling technique, they remain surprisingly constant from one group to another (provided also that the different groups are genuinely comparable).

It is not difficult to deduce a simple working equation which can be applied to data of the kind we are here considering. The general principle is obvious. First of all, to achieve the requisite conversion, we must determine the age-values for (i) the *mean* of the test-deviates for each given age, d_a say ($a = 1, 2, \ldots n$), and (ii), what is more difficult, the *standard deviation* of the individuals tested at each age, namely, σ_a. Then, having due regard to practical requirements, we must select (iii) some standard age to indicate our *zero point* (e.g., age at birth) and (iv) some standard age-difference to define our *unit* of one year.

It will be convenient to begin with a general proof of the equations underlying the procedure proposed. Let us divide the N deviates, which the observed percentages give us for each and every age-group, into two equal sections, say, the first j deviates (in order of increasing size) and the remaining j deviates (where $2j = N$). Now sum the two half columns for every age-group. We

[1] *Loc. cit. sup.*, p. 14. In my later (1921) Report I introduced a slight modification. Having first arranged the tests in a preliminary order of difficulty (by simply summing or averaging the unconverted percentages), it seemed expedient to work with the *differences between the normal deviates* for successive pairs of tests. This enables the investigator to note irregularities which (in the phraseology introduced by Müller) are termed ' inversions of the first and second order ' (6, p. 58). These can often be accounted for by referring back to special circumstances observed during the course of the actual experiments ; and the values causing such irregularities can then be dropped or otherwise adjusted (as in the constant method). Thurstone's ' method of successive intervals ' (unpublished note) appears to use a somewhat similar principle : cf. 27 and refs.

For a fuller account of the procedure, see the passage already quoted in *Mental and Scholastic Tests* (1921, p. 138), which gives a diagram illustrating the results so obtained with the Binet tests (Fig. 21, p. 139 : detailed figures were given in the original memorandum). The items collected for the Graded Reading, Spelling, and Arithmetic tests were calibrated in the same way, in order to secure an evenly graded set of problems for each test in its final form : (5, pp. 270f.).

obtain for the two semi-totals $\sum_{t=1}^{t=j} d_{ta} = S_{a1}$ (say), and $\sum_{t=(j+1)}^{t=N} d_{ta} = S_{a2}$ (say).

(i) *Means.* On taking the *sum* of these two semi-totals, we have, for the mean of the deviates of the ath age-group,

$$\bar{d}_a = \frac{1}{N} \sum_{t=1}^{t=N} d_{ta} = \frac{1}{N} (S_{a1} + S_{a2}) = \frac{1}{N} S_a \text{ (say).} \qquad \text{(viii)}$$

Moreover, using (ii), we also have

$$\bar{d}_a = \frac{1}{N} \Sigma \frac{m_t}{\sigma_a} - \frac{c_a}{\sigma_a} = \frac{1}{\sigma_a} (\bar{m}_t - c_a) . \qquad \text{(ix)}$$

(ii) *Standard Deviations.* On taking the *difference* of the two semi-totals, we have

$$D_a \text{ (say)} = S_{a1} - S_{a2} = \frac{1}{\sigma_a} \left\{ \left(\sum_{t=1}^{t=j} m_t - j c_a \right) - \left(\sum_{t=j+1}^{t=N} m_t - j c_a \right) \right\} = \frac{1}{\sigma_a} M_d, \qquad \text{(x)}$$

where

$$M_d = \sum_{1}^{j} m_t - \sum_{j+1}^{N} m_t.$$

Since by our assumptions M_d must be the same for all ages in the batch, we obtain for any two particular age-groups, α and β say,

$$\frac{D_\beta}{D_\alpha} = \frac{S_{\beta 1} - S_{\beta 2}}{S_{\alpha 1} - S_{\alpha 2}} = \frac{\sigma_\alpha}{\sigma_\beta} = R_{\alpha\beta}, \qquad \text{(xi)}$$

where $R_{\alpha\beta}$ denotes the ratio of the standard deviation of age-group α to that of age-group β. Suppose now that age-group α is the year we choose as the standard year, and β some other year. Then, to express $\bar{d}_\beta$ in terms of the standard year, we must calculate

$$\frac{\bar{d}_\beta \sigma_\beta}{\sigma_\alpha} = \frac{\bar{d}_\beta}{R_{\alpha\beta}} = Q_\beta \text{ (say).} \qquad \text{(xii)}$$

Further,

$$Q_\beta - Q_\alpha = \frac{\bar{d}_\beta}{R_{\alpha\beta}} - \frac{\bar{d}_\alpha}{R_{\alpha\alpha}} = \frac{\sigma_\beta}{\sigma_\alpha} \bar{d}_\beta - \bar{d}_\alpha$$

$$= \frac{\sigma_\beta}{\sigma_\alpha} \cdot \frac{1}{\sigma_\beta} (\bar{m}_t - c_\beta) - \frac{1}{\sigma_\alpha} (\bar{m}_t - c_\alpha) = \frac{1}{\sigma_\alpha} (c_\alpha - c_\beta) . \qquad \text{(xiii)}$$

Now let β be a year adjacent to that which we have chosen as our standard. With the organization obtaining in English schools, age 10—the last age before the pupils are split among different secondary schools—will usually be the best to select as the standard ; and the interval between 9 and 10 years will probably be the best to select to define the unit of a mental year.[1] In that case

$$c_\alpha - c_\beta = 10 - 9 = 1 \text{ ; and } Q_9 - Q_{10} = \frac{1}{\sigma_{10}} (c_{10} - c_9) = \frac{1}{\sigma_{10}} . \qquad \text{(xiv)}$$

Moreover, consider any other age, a say, then, by equation (xiii), $Q_a - Q_{10} = E_a$ (say) will express the deviation of the mean mental age of that age-group (c_a) from the mean mental age of the standard age-group (c_{10}) in terms of the standard deviation of that age-group (σ_{10}).

We now have all that is needed to convert the normal deviates for any test in any age-group to terms of the age and age-difference chosen as standards : for if x_{ta} denote the required converted value of the normal deviate for the tth test and the ath age-group, then in keeping with equation (i) (which expresses the value in terms of absolute years) we have

$$x_{ta} = \frac{\sigma_a}{\sigma_{10}} d_{ta} + \frac{c_a - c_{10}}{\sigma_{10}} = \frac{d_{ta}}{R_a} + E_a, \qquad \text{(xv)}$$

and for the average age-value of any test we at last obtain

$$\bar{m}_t = c_{10} + \sigma_{10} \left(\frac{d_t}{R_t} + E_t \right) \qquad \text{(xvi)}$$

where the subscript t denotes *total* values.

Finally, the accuracy with which the inferred mental ages fit the data from which they have been derived can, if we so desire, be tested by the chi-squared procedure (cf. 5, p. 447).

[1] This is a point at which (as so often in problems of applied psychology) the principles of information and of communication threaten to conflict. By basing our unit on data for two selected age-groups only, we are necessarily losing some of the precision that was gained by starting with a large number of age-groups. But, as we recognized at the outset, every further stage in the process of communicating knowledge is bound to increase ' entropy,' i.e., to lose some of the information still available after the previous stages.

TABLE II. PERCENTAGE PASSING EACH TEST

Test	Age 8·0	Age 9·0	Age 10·0	Age 11·0	Age 12·0	Age 13·0	Total	Average
1	99·96	100·00	100·00	100·00	100·00	100·00	599·96	99·99
2	99·82	99·96	100·00	100·00	100·00	100·00	599·78	99·96
3	99·38	99·85	99·96	100·00	100·00	100·00	599·19	99·86
4	95·22	98·68	99·62	99·88	99·96	100·00	593·36	98·89
5	79·76	93·07	97·72	99·24	99·73	100·00	569·52	94·92
6	66·13	86·67	95·22	98·30	99·38	100·00	545·70	90·95
7	41·37	71·05	87·83	95·22	98·13	100·00	493·60	82·27
8	33·87	64·43	84·13	93·51	97·40	99·97	473·31	78·88
9	20·24	50·00	74·75	88·73	95·22	99·87	428·81	71·47
10	13·33	40·32	67·16	84·38	93·08	99·69	397·96	66·33
11	7·26	28·95	56·63	77·58	89·44	99·18	359·04	59·84
12	3·04	17·71	43·37	67·54	82·44	97·72	311·82	51·97
13	·62	6·93	25·25	50·00	71·05	91·92	245·77	40·96
14	·18	3·22	15·87	38·09	60·91	84·13	202·40	33·73
15	·04	1·32	9·12	27·17	50·00	72·57	160·22	26·70
16	·00	·15	2·28	11·27	28·95	42·07	84·72	14·12
17	·00	·00	·38	3·46	13·33	15·87	33·04	5·51
18	·00	·00	·04	·76	4·78	3·59	9·17	1·53
19	·00	·00	·00	·07	·91	·26	1·24	·21
20	·00	·00	·00	·00	·07	·00	·07	·01

VI. A WORKED EXAMPLE

Practical Computation. The method just described will be easier to understand if we apply it to an illustrative set of data, artificially constructed to yield exact results. To preserve the characteristics of an actual problem, I have based the figures on the results of an empirical investigation.

In attempting to devise a non-verbal test of the matrix type (3 × 3 analogies), Miss Horton and I began by collecting a series of problems which we believed would yield a convenient scale, with two items for each successive year, for children whose mental ages varied from about 5 to 15. After giving the problems to a group of children of a sufficiently wide range, we calculated the percentage of successes scored at each age ; and then determined by graphic interpolation the age-values for all items used. From these we selected a scale of 20 items which we hoped would be equally spaced. Here, in order to secure a simplified example, I have taken the age-values obtained in this pilot survey, and rounded off the figures. From these I have reconstructed the percentages one might expect if there were no errors of measurement or sampling.

The observed percentages, indicating the proportionate numbers of children passing the several tests at each successive age, are shown in Table II. To simplify the arithmetic, the age-groups have been so chosen that no fractions are involved, thus, age 8·0 means children aged $7\frac{1}{2}$ to $8\frac{1}{2}$. I shall suppose that the tests have been applied to all the children in a London borough. With these large numbers, we should be justified in giving percentages to two decimal places ; and we can safely suppose the numbers in the several age-groups to be approximately equal.

Our immediate object will be to treat these percentages as observed proportions, and use them in order to estimate the degree of difficulty presented by each item. If the method proposed is sufficiently trustworthy, the values ultimately reached should be identical with those from which the data were originally derived.

A. Determination of Relative Difficulty. 1. Before embarking on the detailed calculations it will be convenient to *begin by re-arranging the tests in a rough order of difficulty*, as indicated by the total or average percentage for each of the tests : (last 2 columns of Table II).

2. Now *convert each percentage into standard deviation form.* These standardized measures —the normal deviates—are shown in Tables III and VI. In undertaking this, we shall be confronted with an obvious difficulty. With the easiest items as many as 100 per cent. of the older and brighter children pass successfully ; with the hardest all the younger children fail. Now these extreme proportions yield no finite value in standard measure, and must therefore be ignored in the subse-

quent calculations [1] ; and for the reasons given above we shall also do well to reject any very high or very low proportions which seem to be yielding unreliable deviates.

3. Had our percentages been actual figures, they would have contained small irregularities and inconsistencies due to errors. To eliminate or reduce their effect, our plan is to sum or average the corresponding deviates entered in each row. But, if the sums are to be comparable, they must be based on the same set of tests and the same set of age-groups. Now, since we have decided to omit the extreme proportions, we shall not possess data for *all* the tests as applied to *all* the age-groups. We must therefore break up our initial table into workable sections, and *select overlapping batches of tests for which there are reliable data for as many age-groups as possible.* To judge reliability, the simplest method is to inspect any questionable row of deviates, and observe whether the extremely high or low percentages cause the deviates to fluctuate so widely and so erratically that any average based on them would seem almost meaningless. (If the standard deviations of the age-groups differ, that of itself will produce a certain amount of variation : hence it is better to compare the *differences* between the deviates, noting especially any ' inversions,' shown by a change of sign : see above, p. 118, footnote 1.) Here, as the numbers tested are (we have supposed) unusually large, we can safely retain a few rather extreme percentages which otherwise it would be prudent to omit.

It is convenient to select batches containing an *even* number of tests. Here one batch can be taken from age-groups 8 to 12 (Table III) ; these appear to supply pretty regular figures for tests 4 to 15 : a second batch can be taken from age-groups 10 to 13 ; these will yield figures for tests 9 to 18 (Table VI). The remaining tests at either end of the series (tests 1 to 3 and tests 19 and 20) will have to be dealt with separately by methods to be described in a moment.

TABLE III. NORMAL DEVIATES FOR TESTS 1–15

Test	Age 8·0	Age 9·0	Age 10·0	Age 11·0	Age 12·0	Total	Grading	Mental Age
1	−3·333	—	—	—	—	—	−4·000	4·0
2	−2·916	−3·333	—	—	—	—	−3·666	4·5
3	−2·500	−2·963	−3·333	—	—	—	−3·333	5·0
4	−1·666	−2·222	−2·666	−3·030	−3·333	−12·919	−2·666	6·0
5	− ·833	−1·481	−2·000	−2·424	−2·777	− 9·517	−2·000	7·0
6	− ·416	−1·111	−1·666	−2·121	−2·500	− 7·816	−1·666	7·5
7	·208	− ·555	−1·166	−1·666	−2·083	− 5·264	−1·166	8·25
8	·416	− ·370	−1·000	−1·515	−1·944	− 4·413	−1·000	8·5
9	·833	·000	− ·666	−1·212	−1·666	− 2·712	− ·666	9·0
Semi-total ..	−1·458	−5·740	−9·166	−11·969	−14·305	42·641		
10	1·111	·247	− ·444	−1·010	−1·481	−1·578	− ·444	9·33
11	1·458	·555	− ·166	− ·757	−1·250	− ·160	− ·166	9·75
12	1·875	·966	·166	− ·455	− ·972	1·541	·166	10·25
13	2·500	1·481	·666	·000	− ·555	4·093	·666	11·0
14	2·916	1·852	1·000	·303	− ·277	5·793	1·000	11·5
15	3·333	2·222	1·333	·606	− ·000	7·495	1·333	12·0
Semi-total ..	13·194	7·284	2·555	−1·313	−4·537	17·184		

4. Begin with the largest of the batches (Table III), and *sum the deviates in each row.* The totals, taken as they stand, will indicate the relative degree of difficulty of each test within its own batch. If there had been only one batch, and if we were content to estimate merely the *relative* degree of difficulty, then our task would now be finished. But if, as in the present instance, there is more than one batch, then the sums thus obtained from the first batch will not be comparable with those obtained from the rest.

[1] Unless of course we introduce some supplementary device to meet this special difficulty. In the probit method a convenient device has been proposed by Fisher and Yates, which is analogous to that occasionally used in factorial problems, where a similar difficulty arises : see (16), p. 9. For the rougher needs of item-grading it is, in my view, simplest to omit such items or such age-groups, and include them in a second or supplementary batch as described in what follows (see p. 125).

5. Accordingly, we now *apply the same procedure to the second batch, summing the deviates in each row* (Table VI). If there were three or more batches, we should do the same with each in turn.

6. We have now to *convert the sums from each batch to terms of the same arbitrary origin and unit, preferably those of the largest batch.* There are two simple ways of effecting this, (i) a graphic and (ii) an arithmetical. (i) Taking the tests common to the two batches, the beginner usually finds it easier to plot the sums obtained from the largest batch as against the sums obtained from the smaller. (If there are more than two batches, it saves time to use the largest batch for the base-line.) The line of closest fit can then be determined, with fair approximation, by the aid of a transparent ruler ; and the slope and the intercept will furnish the constants for the equation of transformation. (ii) To obtain the equation arithmetically, the quickest method is to divide the data into two sub-sections, and average the sums from each batch, as shown in Table IV.

TABLE IV. TO DETERMINE THE TRANSFORMATION EQUATION

		Sums from Table VI	From Table III	
Tests 9–11 ..	Total ..	−16·788	−4·450	
	Average ..	− 1·483	5·596	
Tests 12–15 ..	Total ..	− 3·185	18·922	
	Average ..	− ·796	4·730	
Given equations..		$a - 5{\cdot}596b = -1{\cdot}483$		(i)
		$a - {\cdot}796b = 4{\cdot}730$		(ii)
On subtracting (ii) from (i) ..		$4{\cdot}800b = 6{\cdot}213$		
		$b = 1{\cdot}2945$		

Substituting for b in either (i) or (ii) we find $a = 5{\cdot}7611$. Hence the equation required is $x = by + a = 1{\cdot}2945y + 5{\cdot}7611$, where $y \equiv$ deviation from smaller batch, and $x \equiv$ corresponding deviation in terms of the origin and unit of the larger batch.

By means of this equation the relative difficulty of each test in Table VI can now be reduced to terms precisely comparable with the figures which express the relative difficulty of the several tests in Table III. With empirical figures the two determinations for the overlapping tests (9 to 15) would doubtless differ slightly ; for these tests, therefore, the two values would have to be averaged.

B. Determination of Mental Age. As we have seen, for composite tests, like the Binet scale or the graded educational tests described in my memorandum, it is convenient to proceed a step further, and convert the arbitrary scale of difficulty into the form of an age-scale, in accordance with the algebraic equations set out above (p. 119). This entails a somewhat complicated calculation. For general purposes the following is perhaps the quickest procedure.

1. As before, begin with the batch containing the largest even number of tests (Table III), and *divide it into two sub-sets with an equal number of tests in each* (here the first 6 and the last 6 tests).

2. Now *sum the figures down each half column.* These sums or semi-totals (S_{a1} and S_{a2}) are shown in Table V, lines 1 and 2. (In this table I have carried the figures to four decimal places in order to assist the reader in checking the calculations, if he wishes : he will note that most of the figures are recurring decimals.)

(i) *To Calculate the Proportional Standard Deviations.*

3. For each column (i.e., each age-group) *subtract the second semi-total from the first.* As we have seen, the difference thus obtained should be proportional to the standard deviation (or the mean deviation) for that group : (Table V, line 3 : ' Difference = D_a ').

4. *Divide each difference by the difference for that particular age which it is proposed to take as a standard* (here age 10 : Table V, line 4 : ' Proportional Difference = R_a ').

5. Using a table of reciprocals (e.g., Barlow's), *find the reciprocals of these quotients.* The values so obtained will show the standard deviations for other ages as multiples of the standard deviation for the standard age (the signs may therefore be dropped : see Table V, line 5, $1/R_a$). With the artificial figures we have been using it is easy to verify that, so far, the procedure gives a satisfactory result : in constructing the original table I had assumed that, to keep the I.Q. constant at least up to age 12, the standard deviations would increase in exact proportion to age ; and this is in accord with the figures just obtained in line 5. Unlike these, the corresponding figure for the *total* does not consist of any simple fraction : thus, if we try to convert the figures in line 3 to vulgar fractions as a check, we find the last to be $\frac{11{\cdot}7222}{59{\cdot}8248} = {\cdot}1959\ldots$

(ii) *To Calculate the Means.*

6. *Add the two semi-totals* (in lines 1 and 2) instead of subtracting them. We then obtain a sum for each column in Table III (tests 4–15 only : see Table V, line 6 : ' Sum = S_a').

7. *Divide these sums by the number* of test-items ($n = 12$). This yields the mean deviate for each column of 12 tests (line 7 : ' Mean = $\bar{d}_a$ ').

8. Now *divide each mean* ($\bar{d}_a$) *by the proportionate value for the standard deviation* (R_a). The quotient or ' Standardized Mean ' (Q_a) is entered in line 8.

9. Finally *subtract this standardized mean* (Q_a) *from the mean obtained for the standard age*, namely, age 10 (Q_{10}). The difference (E_a) expresses the mean of each column as a deviation from the mean of the standard age-group and as a multiple of the standard deviation of that group : (line 9).

(iii) *To Calculate Absolute Values for Means and Standard Deviations.*

10. If we define our unit of mental age in terms of the difference between the standard age and the age next below, *the reciprocal of the difference obtained for these two adjacent years gives the standard deviation of the standard year.* Here, for example, $E_9 = Q_{10} - Q_9 = 0{\cdot}\dot{6} = \frac{1}{\sigma_{10}}$. Hence σ_{10} (the standard deviation of age 10) $= \frac{1}{0{\cdot}\dot{6}} = 1{\cdot}50$ years.

TABLE V. CALCULATION OF MEANS AND STANDARD DEVIATIONS FOR TESTS 1–15

Line	Quantity	Age 8·0	Age 9·0	Age 10·0	Age 11·0	Age 12·0	Total
1.	Semi-totals (Tests 4–9 : S_{a_1})	− 1·4583	− 5·7407	− 9·1666	−11·9696	−14·3055	−42·6410
2.	Semi-totals (Tests 10–15: S_{a_2})	+13·1944	+ 7·2839	+ 2·5555	− 1·3131	− 4·5370	+17·1838
3.	Difference ($S_{a_1} - S_{a_2} = D_a$)	−14·6527	−13·0247	−11·7222	−10·6565	− 9·7685	−59·8248
4.	Prop. Diff. ($D_a \div D_{10} = R_a$)	− 1·2500	− 1·1111	− 1·0000	− ·9090	− ·8333	5·1035
5.	Reciprocal ($1/R_a$)	·8000	·9000	1·0000	1·1000	1·2000	·1959
6.	Sum ($S_{a_1} + S_{a_2} = S_a$)	+11·7361	+ 1·5432	− 6·6111	−13·2828	−18·8425	−25·4572
7.	Mean ($S_a \div 12 = \bar{d}_a$)	+ ·9780	+ ·1286	− ·5509	− 1·1069	− 1·5702	− 2·1214
8.	Stand. Mean ($\bar{d}_a \div R_a = Q_a$)	− ·7824	− ·1158	+ ·5509	+ 1·2176	+ 1·8843	+ ·4156
9.	Dev. of Mean ($Q_a - Q_{10} = E_a$)	− 1·3333	− ·6666	·0000	+ ·6666	+ 1·3333	− ·1353
10.	Dev. of M.A. ($E_a \div E_9 = A_a$)	− 2·0000	− 1·0000	0·0000	+ 1·0000	+ 2·0000	—
11.	Mental Age ($A_a + 10 = M_a$)	8·0000	9·0000	10·0000	11·0000	11·0000	—
12.	Reciprocal of S.D. ($R_a \times E_9 = s_{\bar{d}a}$)	·8333	·7407	·6666	·6060	·5555	—
13.	S.D. of Age-Group ($1/s_{\bar{d}a} = \sigma_a$)	1·20	1·35	1·50	1·65	1·80	

	(i) Age 8	(ii) Age 9	(iii) Age 10	(iv) Age 11	(v) Age 12	(*a*) Total
14. Equations :	$·8\ d_{1a} - 1{\cdot}\dot{3}$	$·9\ d_{1a} - ·\dot{6}$	$1·0\ d_{1a}$	$1·1\ d_{1a} + ·\dot{6}$	$1·2\ d_{1a} + 1{\cdot}\dot{3}$	$·1959\ d_{1a} - ·1353$

11. *Divide the other differences* ($Q_a - Q_{10} = E_a$) *by the selected difference* (i.e., by $Q_{10} - Q_9 = E_9 = 0{\cdot}\dot{6}$) : we obtain the mean mental ages of the age-groups as deviations from the standard age-group : (A_a line 10).

12. *Add the mental age of the standard age-group* (here 10·0). This gives us the absolute *mental* age (M_a) for the other age-groups : (line 11).

13. *Multiply reciprocal of the proportional difference* (R_a) *by the selected difference* (E_9). This will give the *reciprocals of the standard deviations of the age-groups* ($s_{\bar{d}a}$) : (line 12).

14. The reciprocals of these will give the *standard deviations of the age-groups* (σ_a) : (line 13).

(iv) *To Eliminate the Effect of Differing Standard Deviations.*

15. We now have the figures necessary for converting the deviates for any of the age-groups to terms of the age chosen as a standard in accordance with equation (xv). The equations so obtained are given in the bottom line of Table V.

Thèse equations are useful for two main purposes. (*a*) In certain enquiries, where the errors are large and the range is wide, the differences between the standard deviations for the youngest and the older groups may be so great as to introduce an irrelevant differential weight, when the rows are summed to average out the errors : in such cases it may seem better to reduce each column of deviates to terms of the same standard deviation, viz., that of the standard age, and re-calculate the sums.[1] (*b*) But the chief use of the equations will be to evaluate outlying figures from test-items not included in the selected batches dealt with by the methods of Tables V and VI. For example, for test 1 we may take the deviate-value — 3·3333 from age-group 8·0 in Table III ; the value for $1/R_a$ we can take from line 5 in Table V, and M_a from line 9 in Table V. We thus obtain

$$\cdot 8 \times (-\ 3\cdot 3333) - 1\cdot 3333 = -\ 4\cdot 0000.$$

16. For the tests in the selected batches we shall use the equation given at the end of line 14, and apply it to the summed deviates. Thus for test 4 we have

$$\cdot 1959 \times (-\ 12\cdot 919) - \cdot 1353 = -\ 2\cdot 6666.$$

The gradings so obtained have been appended in the last column but one of Table III. These gradings treat an ideal test of median difficulty for age 10 as marking the zero-point and the standard deviation of age 10 as marking the unit.

17. Our ultimate aim, however, is to express the final value for each test in terms of a mental age. For this purpose the necessary transformation is obvious, viz.,

$$M.A. = \text{Age of Standard Group} + \frac{\text{Grading}}{S.D. \text{ of Standard Group}},$$

that is, $M.A. = 10\cdot 0 + \frac{G}{0\cdot 6}$, where $M.A.$ and G are now written for $\bar{m}_t$ and x_{ta}.

I have taken this final calculation in two steps, so that the beginner may appreciate the underlying principles. Using equation (xvi) (p. 119) the practical computer will construct an equation which will enable him to convert the totals directly to mental ages without calculating the intermediate ' gradings.' Taking the values from lines 9, 5, and 12 of Table VI, we have

$$M.A. = 10\cdot 0 + \frac{1}{0\cdot 6}(\cdot 1959\ d_t - \cdot 1353)$$

$$= 9\cdot 7972 + \cdot 2938\ d_t,$$

where d_t denotes the total given in Table III.

Second Batch. The detailed calculations for the second batch of overlapping tests (9 to 18) are shown in Table VI. If the totals have already been converted to terms of the origin and standard deviation of the totals in Table III, the same equation can be used for obtaining the corresponding mental ages. As a rule, however, it will be quicker to construct an equation for transforming the original totals (Table VI) direct to mental ages. Using the same procedure as before, we obtain for this batch $M.A. = 11\cdot 4904 + \cdot 3805\ d_t$. Tests 19 and 20 will of course have to be dealt with by the method described above in step 15.

[1] The only case in which to my knowledge it has been used was in an unpublished study by Mr. Walton, which showed that with data from the Binet scale only the second decimal place was affected ; and in point of fact such corrections will rarely be necessary : nobody appears to require so refined a degree of accuracy in assessing item difficulty.

In describing the scaling methods used by those who prefer a ' P.E. scale ' Garrett does not explain how results obtained from more than one group are to be combined (18, pp. 146–59). Actually American investigators who have adopted the procedure outlined by Garrett do not, as a rule, test separate age-groups, but separate grades. On this ground they apparently assume that it will be unnecessary to take into account the differing standard deviations in the grades. As a consequence, to estimate the differences between the means for consecutive grades they merely calculate the grade-differences for the several tests and average those. The reason commonly advanced for their assumption is that, in re-grouping children into grades, the teachers would naturally keep the range and variability in the higher grades much the same as in the lower. The figures actually reported, however, seldom bear out this view. Usually the differences between the normal deviates tend to increase with the difficulty of the tests. This is sometimes explained by claiming that the brighter children are best discriminated by harder tests. In point of fact, it is a direct consequence of the increase in the standard deviation. In the example here taken the difference between ages 8 and 9 for test 4 (for instance) is — 1·6666 — (— 2·2222) = ·5555 ; for test 15 it is 3·3333 — 2·2222 = 1·1111 —just twice as much. It is therefore quite illegitimate to *average* such differences. That would be equivalent to taking the difference between the averages of the age-groups (given in line 7 of Table V) just as they stand. Before taking the difference we must standardize each average by dividing by R_a, as described above.

TABLE VI. NORMAL DEVIATES FOR TESTS 9–18

Test	Age 10·0	Age 11·0	Age 12·0	Age 13·0	Total	Grading	Mental Age
9	— ·666	—1·212	—1·666	—3·000	—6·545	— ·666	9·00
10	— ·444	1·010	—1·481	—2·733	—5·669	— ·444	9·33
11	— ·166	— ·757	—1·250	—2·400	—4·574	— ·166	9·75
12	·166	— ·454	— ·972	—2·000	—3·260	·166	10·25
13	·666	·000	— ·555	—1·400	—1·289	·666	11·00
1. Semi-total	— ·444	—3·434	—5·925	—11·533	—21·338		
14	1·000	·303	— ·277	—1·000	·025	1·000	11·5
15	1·333	·606	·000	— ·600	1·339	1·333	12·0
16	2·000	1·212	·555	·200	3·968	2·000	13·0
17	2·666	1·818	1·111	1·000	6·596	2·666	14·0
18	3·333	2·424	1·666	1·800	9·224	3·333	15·0
2. Semi-total	10·333	6·363	3·055	1·400	21·152		
3. Difference	—10·7777	—9·7979	—8·981	—12·9333	—42·4906		
4. Prop. Difference ..	— 1·0000	— ·9090	— ·8333	—1·2000	— 3·9425		
5. Reciprocal ..	1·0000	1·1000	1·2000	·8333	·2536		
9. Dev. of Mean ..	·0000	·6666	1·3333	1·8333	·9936		
11. Mental Age ..	10·0	11·0	12·0	12·75	—		
12. Reciprocal of S D...	·6666	·6060	·5555	·8000	—		
13. S.D. of Age-Group	1·50	1·65	1·85	1·25			

	(iii) Age 10	(iv) Age 11	(v) Age 12	(vi) Age 13	(*b*) Total
14. Equations :	$1{\cdot}0d_{ta}$	$1{\cdot}1\ d_{ta}+{\cdot}6$	$1{\cdot}2\ d_{ta}+1{\cdot}\dot{3}$	$1{\cdot}275\ d_{ta}+1{\cdot}\dot{8}\dot{3}$	$\cdot 2536\ d_{ta}+\cdot 9936$

Note on the Final Evaluation. The mental ages for the tests in the two main batches are given in the last columns of Tables III and VI. From the values thus reached for the whole series we see that the distribution of degrees of difficulty throughout the scale is distinctly uneven. There is a bunching near the middle of the range ; and the increments in difficulty are by no means equal from one item to the next. The interval 5·0 to 7·0 years (inclusive) contains only three test-items (one per year) ; the intervals 7·0 to 9·0 years and 9·0 to 11·0 years each contain five : in point of fact the original intention was to proceed by half-yearly intervals from 4·5 to 14·0 years. Accordingly, with a scale like this, which contains only 20 items, it would be strongly advisable to amend these defects before putting it to practical use.

Here, however, we are concerned rather with the efficacy of the statistical procedure. The mental ages reached by the calculations prescribed prove to be identical with those from which the percentages in Table II were originally constructed. For artificial data that contain no error, therefore, the method is perfectly exact. With empirical data errors will always be present. The procedure proposed assumes that, on averaging or summing the rows and the half-columns, these errors will neutralize one another. Further, owing to such errors, the mental ages obtained for the tests included in *both* the overlapping batches (9 to 15) would not be precisely the same in the two cases ; and, as noted above, they too will have to be averaged. But practical application to actual data, obtained both from mental and from scholastic tests, indicates that, so long as we restrict ourselves to item-deviates that involve no great scope for error, the method of summing and averaging sufficiently eliminates the inaccuracies involved.

The method of sets of assigning items to appropriate ages will be clear from what has been said above. If we adopt the principle recommended in my memorandum for scaling the tests of reasoning and scholastic attainments (see p. 108 above), then (assuming our age-values have been obtained to two decimal places) items having an age-value between (say)

8·005 and 9·005 years would be assigned to age 9. This means that " theoretically a child on its ninth birthday should pass all the tests above the heading ' 9 years ' and none of the tests below that heading." As noted above, however, it would be wise to make a final check by calculating the average mental ages thus awarded to children of specified chronological ages to ensure that the method of marking prescribed does give the requisite results in actual practice.

VII. ABSOLUTE SCALING

In psychology an ' absolute scale ' may be defined as a ratio scale in which both zero-point and unit are independent of the particular test employed and of the particular group or groups that have been tested.[1] Thurstone's attempt to produce such a scale is well known (7). His aim, however, is somewhat different from mine. My object was chiefly practical—to facilitate the construction of graded tests that could serve as trustworthy instruments of measurement. Thurstone's interest, as he says, lies rather in " theoretical problems " : he is not concerned with " tests constructed merely for administrative purposes," but with more abstract issues—the curve of mental growth, the rate of retardation, and the change in variability from age to age or grade to grade.[2] For purposes such as these, so he contends, it is essential to devise some method of scaling which shall be " independent of the units represented by the raw scores."

In his first article he describes *two* procedures for re-scaling the Binet tests with the aid of the percentages I had obtained with London children.

1. For the first he simply takes two percentages for each test (those on either side of 50 per cent.), and applies what is really the ordinary formula for the median (see above, p. 111). In this way he seeks to express the difficulty of each test in terms of " years at par." The results are given in his final table (*loc. cit.*, Table V, p. 451). As we have already seen, the order so reached corresponds almost exactly with that which I had got by summing all the percentages available for each test. And, indeed, as he remarks, this part of his analysis is " similar to methods already in use."

2. In order to express the difficulty of each test in ' absolute ' terms he proposes a more elaborate procedure. He begins by taking the mean of age $3\frac{1}{2}$ as an arbitrary zero and the standard deviation for the same age as his arbitrary unit. The mean and standard deviations of the next age-group ($4\frac{1}{2}$) are then expressed in terms of the previous age-group ($3\frac{1}{2}$) ; then the third age-group ($5\frac{1}{2}$) is reduced to terms of the second, and so, indirectly, to that of the first ; and so on. When he turns to estimate the difficulty of the individual tests (a matter which for him is mainly incidental), he again takes *two* percentages only for each one, and averages the two results. He gives no table of the results achieved by these more arduous calculations, but merely presents them in the form of a linear diagram. From this it would seem that the chief effect of his chain-like procedure is gradually to increase the spacing between the several tests and ages : thus between ages $11\frac{1}{2}$ and $12\frac{1}{2}$ the interval is apparently almost double what it is between ages $3\frac{1}{2}$ and $4\frac{1}{2}$. Apart from this, the general order and the relative spacing, as well as the marked bunching between ages 6 and 7 and 11 and 12, and the wide gaps elsewhere, appear to be almost exactly the same as those shown in my own diagram (5, Fig. 21, p. 151).

It is Thurstone's second procedure that is described and recommended by Guilford (10, pp. 440-3). Holzinger (8) criticizes it in some detail, principally on the ground that the basic assumptions are unfulfilled and that the new and over-elaborate calcu-

[1] Cf. Warren, *Dictionary of Psychology* (1934), *s.v.* The term applies not merely to the scaling of tests, but also to the measurement of individual characteristics ; and not only to tests where the responses are of a right-or-wrong type, but also to those yielding continuously graded measurements. Both the conception and the phrase seem to have been taken from Kelvin's ' absolute scale of temperature.' But, as Miss Goodenough remarks, in psychology any such description is " more hopeful than accurate " (*Mental Testing*, 1950, p. 155).

By a ' ratio scale,' as distinguished from an ' ordinal scale ' (ranking) or an ' interval scale,' I mean one for which there are practicable operations for determining not only (i) the order of the items, but also the equality (ii) of items, (iii) of intervals, and (iv) of ratios. Cf. Stevens, S. S., ' On the Theory of Scales of Measurement,' *Science*, CIII, 1946, pp. 677f.

[2] *J. Educ. Psychol.*, XIX, 1928, p. 124 (Reply to Professor Holzinger).

lations are unnecessary. Thurstone replies by calling attention to the defects of the procedures given in "the current text-books on educational measurement," such as those of McCall and Trabue or Garrett's text-book on *Statistics in Psychology and Education* ("the P.E. method of scale construction"). These, he points out, make the "fallacious assumption" that "the dispersion remains constant for all age- and grade-groups."

The criticisms that I myself should be inclined to urge are the following:

(i) First, the calculations are extremely onerous: in Thurstone's article two pages in small type are required to illustrate the calculation for the relation between the means and the standard deviations for one pair of ages only; the complete evaluation involves similar calculations for ten such pairs. When we remember the kind of data generally available and the doubtful applicability of the assumptions made, any slight gain in rigour that such computations could confer will hardly justify the additional labour. (ii) When (as with the pair taken by Thurstone) the same tests furnish definite figures for all the age-groups with which we are concerned, then (as I have tried to show) it is unnecessary to calculate the standard deviations. On the other hand, when the number of age-groups is so large that different sets of tests are needed, then the change in standard deviations must certainly be allowed for. But Thurstone's method of achieving this seems scarcely satisfactory. Instead of calculating an average (weighted or unweighted) for all the comparable groups, he works, as we have seen, with consecutive pairs. And evidently the errors thus introduced are bound to be cumulative. (iii) Moreover, to determine the relative value of the standard deviations, Thurstone squares the normal deviates. But the larger the normal deviate, the greater its liability to error; and this will inevitably be magnified by squaring.[1] That difficulty is avoided by my own simpler method; and on these grounds I still consider it to be fully justified, at least for the general purposes for which it was originally devised. (iv) When it comes to "finding the scale-values of the test-items," then, as we have seen, he simply uses "the two proportions for each item that are nearest to ·50, one on either side." Like his first method, this seems to ignore a large amount of the available information. (v) Thurstone's formulæ explicitly assume (7, p. 439) that the correlation between the normal deviates for comparable items belonging to two consecutive age-groups is 1·00, or, as he puts it elsewhere, that "the plot [of the two sets of deviates] is linear." This is the point challenged by Holzinger. Holzinger finds that the variance of the discrepancies between the two scale-values is ·088 and its p.e. $\pm$ ·012, and hence infers that "the variance could not have arisen from the fluctuations of random sampling." Thurstone's reply is simply: "Look at the diagrams." Although I readily accept Thurstone's inference as sufficiently correct for practical purposes, the point is scarcely one to be decided by simple inspection.

In my view the main points of conflict could be largely reconciled if we revert to the standpoint suggested at the outset, namely, that of factor analysis. As stated in my original memorandum on the Binet scale, there are undoubtedly numerous irregularities from year to year which have clearly assignable causes, other than the intrinsic difficulties of the test: I do not hold for one moment that *all* the significant variance can be accounted for by the general factor of test-difficulty. On the other hand, this does not to my mind imply that "the scaling of individual items is ordinarily not worth while at all."[2] Rather, in any intensive study, we should attempt to

[1] In a later article (9, pp. 441–53), he partly meets this objection by weighting the normal deviates with what are in effect the Urban-Müller weights. But he still compares his groups by taking them in consecutive pairs. In this paper, however, his ultimate aim is to examine the means and the variabilities of successive school grades in a test of arithmetic. "Our object," he says, "is to represent the overlapping grade groups . . . on a single scale of test ability." Thus, he is no longer concerned with scaling test-items according to their difficulty, and gives no table or graph for this purpose. Moreover, the introduction of the differential weighting renders the calculations more cumbersome still. With my method practically the same results can be reached very simply, as follows. His series of test-items can be divided into two halves, each half containing 13 items—Nos. 2–14 and 15–35. (I omit No. 1 as it appears to contain a misprint; Thurstone has omitted several harder items, for grounds stated in his article.) The semi-totals are then $-$ 11·85 and $+$ 17·59 for grade ii, and $-$ 19·31 and $+$ 5·76 for grade iii. Taking differences, the ratio of the standard deviations is at once found to be 25·07 $\div$ 29·44 = ·85, and of the means to be ·1885 and $-$ ·5259. The final equation is therefore $\cdot 85x_1 + \cdot 69$. Thurstone's equation is $\cdot 84x_1 + \cdot 70$, almost exactly the same figures. An iterative procedure has been developed by Flanagan (12). But he, too, is interested in assessing the groups and their individual members rather than the difficulty of test-items, and employs only the median score of the several distributions.

[2] Holzinger, K. J., 'Reply to Professor Thurstone,' *loc. cit. sup.*, p. 124.

estimate *the proportionate amount* which the factor of test-difficulty contributes to the total variance ; and then, if the residue is still found to be significant, consider what irrelevant subsidiary factors are influencing the results observed.

VIII. SUMMARY AND CONCLUSIONS

1. Current criticisms of the practice of mental testing arise partly from alleged defects in its theoretical foundations and partly from undoubted shortcomings in the results obtained. Such criticisms can best be met by examining more closely the practical requirements which the psychologist seeks to fulfil : recent developments in the theory of communication and information, based on practical needs, may be usefully adopted to build up sounder working principles for improving test-procedures in applied psychology. And it is argued that far more efficient results could be obtained if fuller consideration along these lines were given to methods of test-construction, and particularly to the problem of scaling the component items in accordance with their difficulty.

2. Current text-books dealing with psychometrics pay little attention to such considerations ; and consequently most of the published tests appear to have been compiled without proper empirical checks. The majority consist of items marked on a pass-or-fail, right-or-wrong basis ; but these items have usually been selected and arranged by the rough-and-ready procedures used by Binet and his earlier followers.

3. Binet's final procedure consisted in grading the test-items in order of difficulty and then assigning successive sets to an appropriate age on the basis of a percentage criterion. This method appears sound so far as it goes. But, even in the latest revisions of his scale, there seems no general agreement about what particular percentage should be taken as the criterion, and little or no attempt to ensure that the scale should show equal increments of difficulty.

4. If we accept the principle that the most efficient procedures are those which make a maximum use of the available information, then the best method of estimating the difficulty of the test-items would be to convert the percentages of failures at each age into normal deviates, and calculate the factor-measurements to be obtained by correlating the weighted figures for the several age-groups. But for ordinary purposes simpler methods may be devised.

5. Those which have been or might be used can be classified as follows :

A. Methods Based directly on Observed Percentages

1. Median calculated by Linear Interpolation.
2. Median obtained by Graphical Interpolation.
3. Mean calculated by Summing Percentages.

B. Methods Based on Normal Deviates

4. Median of Best-fitting Ogive : (Constant or Probit Method).
5. Test-difficulty as a General Factor.
6. Test-difficulty determined by Simple Summation.
7. Test-difficulty expressed in Mental Years.
8. Absolute Scaling.

6. Each of these has been described and examined in turn ; and it is concluded that, for nearly all practical requirements, the most satisfactory method will be to add or average the normal deviates, omitting the extreme values on account of their unreliability, and making due allowance (when necessary) for differences in the variability of successive age-groups—the method used for the 1921 London Revision of the Binet scale. The results can be readily expressed in terms of mental age if desired. Thurstone's proposal to substitute a method of absolute scaling appears difficult to justify either on theoretical or on practical grounds.

7. For the procedure proposed, detailed working instructions are given and illustrated by a worked example.

REFERENCES

1. Binet, A., and Simon, T. (1908). ' Le developpement de l'intelligence chez les enfants,' *Ann. Psychol.*, XIV, 1–94.
2. Binet, A. (1911). ' Nouvelles recherches sur les mesure du niveau intellectual chez les enfants d'école,' *Ann. Psychol.*, XVII, 145–201.
3. Burt, C. (1914). ' The measurement of intelligence by the Binet-Simon Tests,' *Eugen. Rev.*, VI, 36–50, 140–52.
4. Burt, C. (1915). ' The standardization of the Binet-Simon tests for London schools.' *L.C.C. Reports.*
5. Burt, C. (1921). *Mental and Scholastic Tests.* London : P. S. King (Third edition, Staples Press, 1949).
6. Brown, W., and Thomson, G. (1925). *The Essentials of Mental Measurement.* Cambridge : University Press.
7. Thurstone, L. L. (1925). ' A method of scaling psychological and educational tests,' *J. Educ. Psychol.*, XVI, 433–45.
8. Holzinger, K. J. (1928). ' Some comments on Professor Thurstone's method of determining the scale values of test items,' *J. Educ. Psychol.*, XIX, 112–17.
9. Thurstone, L. L. (1928). ' Scale construction with weighted observations,' *J. Educ. Psychol.*, XIX, 441–53.
10. Guilford, J. P. (1936). *Psychometric Methods.* New York : McGraw-Hill Book Co.
11. Yule, G. U., and Kendall, M. G. (1937). *Introduction to the Theory of Statistics* (11th ed. : 1st ed., Yule, G. U., 1910). London : Griffin.
12. Flanagan, J. C. (1939). *Scaled Scores.* New York : Co-operative Test Service.
13. Woodworth, R. S. (1939). *Experimental Psychology.* London : Methuen.
14. Richardson, M. W. (1941). ' The logic of age-scales,' *Educ. Psychol. Measurement*, I, 25–37.
15. McNemar, Q. (1942). *The Revision of the Stanford-Binet Scale.* New York : Houghton Mifflin Co.
16. Fisher, R. A., and Yates, F. (1943). *Statistical Tables.* Edinburgh : Oliver and Boyd.
17. Finney, D. J. (1947). *Probit Analysis: A Statistical Treatment of the Sigmoid Response Curve.* Cambridge : University Press.
18. Garrett, H. E. (1947). *Statistics in Psychology and Education.* New York : Longmans, Green & Co.
19. Wald, A. (1947). *Sequential Analysis.* New York : John Wiley and Sons.
20. Wiener, N. (1948). *Cybernetics.* Paris : Hermann et Cie.
21. Shannon, C. E., and Weaver, W. (1949). *The Mathematical Theory of Communication.* Urbana : University of Illinois Press.
22. Wald, A. (1949). ' Statistical decision functions.' *Ann. Math. Statist.*, XX, 165–205.
23. Burt, C. (1950). ' The theory of mental testing,' *Third Yearbook of Education*, Evans Brothers.
24. Gulliksen, H. (1950). *Theory of Mental Tests.* New York : John Wiley and Sons.
25. Mackay, D. M. (1950). ' Quantal aspects of scientific information,' *Phil. Mag.*, 7th Ser., XLI, 289–97.
26. Shannon, C. E. (1950). ' Programming a computer for playing chess.' *Phil. Mag.*, 7th Ser., XLI, 256–75.
27. Edwards, A. L. (1951). *Psychological Scaling by means of Successive Intervals.* Chicago : Psychometric Laboratory.

CORRESPONDENCE

To the Editors, British Journal of Psychology, Statistical Section

SIRS,—In discussing the calculation of multiple correlations (this *Journal*, IV, p. 53), Professor Burt states that " it is inaccurate to apply a multiple correlation, calculated for a given sample, to some other sample." He quotes a correction formula, which to the non-statistical psychologist looks somewhat paradoxical, since $\bar{R}^2$ (the probable value of the multiple correlation) appears to be reduced *more* rapidly when *more* predictor tests are used. Guilford refers to this possibility (*Psychometric Methods*, p. 426), but suggests neither this nor any other formula. However, the very fact of ' shrinkage ' is disquieting to the practical psychologist, who, following the precedent set in so many published papers and text-book examples, usually fancies his task is virtually over when he has calculated the appropriate weights or regressions, together with the multiple correlations derived from them in the ordinary way. Could Professor Burt give some concrete information on the amount of shrinkage we should expect in actual work, say, for example, in weighting the various tests set in a scholarship examination of the ordinary type, and could he also indicate suitable references for psychologists engaged in such work?

J. R. DAWSON

Reply. In general, with a well-selected battery of tests, the absolute value of the multiple correlation tends to increase with the number of tests, though usually at a diminishing rate. The so-called ' shrinkage coefficient ' merely states that, when the implications of such a correlation are considered for a second sample, the value calculated from the first will be appreciably reduced, and that the amount of reduction will (in effect) be greater for the larger values derived from a larger number of tests.

In his more recent publication (*Fundamental Statistics in Psychology and Education*, 1950, p. 434), Guilford quotes and illustrates the shrinkage formula, without, however, indicating how it is derived. Adopting his notation, let σ_1 denote the standard deviation of the criterion (x_1) as estimated from the given sample, and let $\sigma_{1\cdot 23\ldots m}$ denote the ' standard error of an estimate from a multiple prediction.' We then have for the estimate of ' the standard error in the population when corrected for bias '

$$_c\sigma^2_{1\cdot 23\ldots m} = \sigma_1^2(1 - R^2)\frac{N-1}{N-m-1}. \qquad \text{(i)}$$

Since $\bar{R}^2$ (the corrected value) $= 1 - \dfrac{\sigma_1^2}{{}_c\sigma^2_{1\cdot 23\ldots m}}$, we at once obtain the correction formula proposed. (I have taken the degrees of freedom as given in Fisher's version of the usual test, *Statistical Methods for Research Workers*, Sect. 47, p. 247 : Guilford adopts the older usage which gives the denominator as $N - m$.) The above equation (i) leads to a rough and suggestive criterion, which may be usefully employed, before we embark on the more elaborate calculations for testing the significance of R : the estimated error variance ($_c\sigma^2_{1\cdot 23\ldots m}$) should obviously be smaller than the estimated criterion variance (σ_1^2) ; and on substituting the expression given by (i), we at once see that R should be greater than $\sqrt{\dfrac{m}{N-1}}$. This is equivalent to asking whether, with a sample of 101 children, R (or r) obtained from 1 test is over 0·1, from 4 tests over 0·2, from 9 tests over 0·3. Now, if the 9 tests had been applied to 10 children only, we could then evidently find weights which would enable us to ' predict ' the (given) criterion-measurements for the 10 children *exactly* : (9 tests, not 10, are sufficient, because it is assumed that the predicted measurements will be in deviation form, so that the measurement for the 10th child can be deduced from the fact that the sum of all the measurements must add up to zero). In such a case, it will be seen, we have in effect increased

the number of our tests to the maximum required, and with the *given* sample the multiple correlation will be perfect. But plainly we must anticipate a very large reduction in this correlation coefficient, when we apply our regression equation to *another* sample.

This brings out what, in my view, is a most important caution for the practical worker. If we now re-calculate our regressions so as to find the best weights for the *second* sample, the figures we obtain will evidently not be the same as those previously derived from the first. Let me illustrate this by an actual example from the field Mr. Dawson has in mind, described in an earlier publication. With a group of 257 scholarship children the multiple correlation between four selection (or 'predictor') tests and an independent criterion amounted to 0·66 with the 'best' weights (i.e., the partial regressions). With a second batch of children, tested in the following year, the multiple correlation was 0·68 when the 'best' weights were calculated afresh from the sample. Thus the correlation remained much the same, when new weights were used; but these new weights differed appreciably from the old. However, when the old regressions (i.e., those found with the first sample) were used for the new sample, the correlation sank to 0·55; and, when the tests were given equal weights, the correlation was only slightly less, namely, 0·53. If, however, the best three tests were chosen and still given equal weights, it rose to 0·57. The conclusion drawn was that, for rough practical purposes, it is often unnecessary to use differential weights, such as those obtained by the lengthy calculations required for a regression equation: "it may be sufficient merely to take a straight sum or simple average of the marks" (*Brit. J. Psych., Gen. Sect.*, XXXIV, 1943, p. 19).

Where more precise figures are needed, it is evidently essential to try out the weighting scheme, obtained from a first sample A, on at least one other sample, B. British writers usually call this process 're-validation.' Since the time order is irrelevant, we can also calculate the regressions from the second sample, B, and then try them with the first sample A. This double check is called 'cross-validation': (American writers seem often to use the latter term in the former sense). However, as I have contended in earlier papers, two samples may not be enough: a sequential procedure, in which successive groups are tested and regressions repeatedly re-calculated, seems indispensable before a really satisfactory procedure can be achieved. Such a task is scarcely practicable for a single research-student, however competent. Here lies the value (as the policy of the London County Council has shown) of appointing a permanent psychologist, engaged largely on research. But I am very doubtful about the value of introducing weights into scholarship examinations or the like. My advice to the practical psychologist would therefore be this: use the correlational analysis to determine what tests have a low validity or overlap with the rest, drop those with low weights, and then give the selected tests an *equal* weight (e.g., by standardizing the marks or simply averaging ranks [1]).

As regards references, an admirable paper, describing tests for the hypothesis that in the different groups tested the regression planes are the same, has been recently published by Gulliksen and Wilks ('Regression Tests for Several Samples,' *Psychometrika*, XV, 1950, pp. 91–114). Where only two small samples are used, Hotelling's generalization of student's ratio t (the so-called T-test) can be adopted: (for this, see Kendall, M., *Advanced Theory of Statistics*, p. 338 and refs.). For more general discussions the reader may consult Mosier, C., 'Problems and Designs of Cross-validation,' *Educ. Psychol. Meas.*, XI, 1951, pp. 5–11, and Katzell, R. A., 'Cross-validation in Item Analysis,' *ibid.*, pp. 16–22.

CYRIL BURT

SIRS,—In the volume of essays published in honour of Professor Piéron (*l'Année psychologique*, 1951) Professor Thurstone begins his paper on 'Factor Analysis as a Scientific Method' by saying: "Factor analysis was started by the British psychologist Charles Spearman in a remarkable paper published in 1904." He then goes on to illustrate his own views by referring to the factor analysis of a "matrix of anthropometric measurements for a population of adult men." He concludes that "a principal axis solution is indicated"; and observes that, as soon as improved computing machines become more generally available, some such procedure (with perhaps an iterative working method) will be used almost universally: several of his own "approximation procedures," he adds, "will then be obsolete."

[1] This was the procedure adopted in my earliest investigations (*Brit. J. Psych.*, III, 1909, p. 158), and usually gave satisfactory results in examinations carried out for the L.C.C. Dr. Banks tells me that, with figures for the yield of ten main crops, obtained at 5-year intervals (*v.* Kendall, M. G., *J. Roy. Stat. Soc.*, CII, pp. 21–48) the saturations and regressions are very different from one year to the next, but the multiple correlations obtained by simply averaging ranks for the four or five most typical crops (as judged from the first sample) remained well over 0·9. Thus the rule seems to apply in other fields besides that of mental testing. (See also this *Journal*, III, pp. 105–25; 'The Influence of Differential Weighting.')

Now, is not the factorization of a " matrix of anthropometric measurements " precisely the problem which Galton had propounded and to which Karl Pearson gave a solution in papers published as early as 1901 and 1902? (see *Phil. Mag.*, 6th Ser., II, p. 559; and *Biometrika*, I, p. 177). Since Spearman describes his own 1904 papers as largely based on the work and methods of the Galton-Pearson school, should not the credit go rather to Galton for seeing the problem and to Pearson for solving it? Further, I can find nothing in either of Spearman's 1904 papers which refers to factor analysis as a method or even to a ' factor ' in Professor Thurstone's sense. The small table of correlations which he refers to as appearing in Spearman's paper is *not* (as he says) a table of ' tests ' but a table of examination marks and *one* test with children of different ages. Moreover, as Thurstone himself points out, Spearman relied solely on ' inspection ' to prove what was in effect Galton's hypothesis of a ' general ability.' This is hardly an improvement on Pearson's procedure; and certainly, however much criticism and controversy it may have evoked, is surely not " the starting-point of factor analysis." Is there no member of the British Psychological Association whose memory goes back to these early years who can assign, correct, or confirm this piece of history? I am sure that many who are giving elementary instruction would agree that it is a point that urgently needs clearing up.

C. Leclère

[*Editorial Note.*—The earliest publication suggesting the method of principal axes as a solution to Galton's problem is a paper by Edgeworth in *J. Roy. Stat. Soc.*, LIX, 1896, pp. 534–9. Spearman's 1904 paper (i.e., his article on 'General Intelligence Objectively Determined and Measured,' *Amer. J. Psychol.*, XV, pp. 51–84) was the starting-point, not of factor analysis as such (he made no such claim), but of his two-factor theory. This was later developed as a simplified alternative to Pearson's more elaborate procedure—a simplification rendered possible, in Spearman's view, by the absence of any *third* kind of factor in *mental* measurements.

M. Leclère will find a fuller answer to his questions in this *Journal*, II, pp. 151f.]

Vol. IV
Part II

The British Journal of Psychology,
Statistical Section

June,
1951

NOTE ON NORMAL TRANSFORMATIONS OF RANKS

By A. S. C. EHRENBERG

Statistical Laboratory, University of Cambridge [1]

Fisher and Yates (1, Table XX) give a table of 'normal' transformations for ranked data (sample sizes 2 to 50) to enable the user to apply analysis of variance and correlation tests. The transformation consists in giving n ranked objects the expected values of an ordered sample of n readings (the quantiles) from a normal population with zero mean and unit variance. The use of this technique continues to be advocated: it appears, for example, in Quenouille's recent text-book (2, p. 175); but no theoretical basis for it seems to have been given in print.

If the original ranks are set out in the form of integers 1 to n, the transformation to normal quantiles clearly represents a non-linear change of scale. This aspect is mentioned by Bliss *et al.* (3), who say that the 'normalizing transformation' corrects the tendency for extreme items to be ranked more easily. The proposed correction, like any other fixed change of scale, is rather arbitrary, even when such a tendency exists; but it is quite out of place if the objects are essentially identical (a common null hypothesis), or if, for example, each item belongs to one of two classes within each of which variations are relatively small.

The transformed ranks are obviously not a random sample from a normal population, even on the null hypothesis that the ranks are given at random. The usual deductions of the distributions of the standard product-moment correlation coefficient and analysis of variance tests do not therefore apply; and any proposed analysis of transformed ranks would have to be proved valid *ab initio*.

The correlation coefficient need no longer be discussed in this context since more directly applicable techniques exist (cf. 6). As regards the analysis of variance, the customary computations with data composed of m rankings of n objects appear to be rather artificial. The 'between rankings' sum of squares is identically zero, and the 'between objects' and 'residual' sums of squares are not independent, since they add up to the 'total' sum of squares which is a constant for given m and n. When $m(n-1)$ is not too small, however, it has been shown by Pitman (4) that on the null hypothesis of random ranking the ratio of the 'between objects' and the 'residual' mean squares is distributed approximately in accordance with the F-distribution (the variance ratio $F = e^{2z}$, where z is used as in Fisher's notation), with

$$\nu_1 = n - 1 - \frac{2}{m} \text{ and } \nu_2 = (m-1)\,\nu_1$$

degrees of freedom. (When m and n are both small, the tables of the exact distribution given by Kendall and Babington Smith (5) or Kendall (6) can be used; Kendall also discusses the effects of ties.) The approximation to the F-distribution follows because the second moments of all the m rankings are equal, and it is shown that higher moments can be ignored. Since the distributions are independent of the variance in each ranking, the test actually applies equally well to transformed ranks, but appears to be in no way better.

This approximate test for random ranking is closely related to one based on a simpler concept, namely, the Coefficient of Concordance W, the variance ratio being equal to $W/(1-W)$, (5, 6). An approximate one-parameter distribution for the expression $m(n-1)W$ is also given by Friedman (7) as χ^2 with $(n-1)$ degrees of freedom, when $m(n-1) > 7$; this illustrates perhaps that only the 'between objects' sum of squares can vary independently.

[1] This note has been written whilst carrying out work on behalf of the Department of Scientific and Industrial Research (Food Investigation Organization).

It may be of some interest that an exact test based on the F-distribution for the null hypothesis can be obtained by scoring the n objects in each ranking with the ordered values of an actual (not an average) random sample of size n from a standard normal distribution. Such samples can be obtained from Mahalanobis *et al.* (8) or Wold (9), or by calculation from tables of random numbers. On the null hypothesis all the sums of squares in an analysis of variance of the transformed data are distributed as χ^2 with the usual degrees of freedom. Clearly, the variance ratio test of the ' between rankings ' mean square against the residual is quite irrelevant even if it turns out significant ; but it does account for the appropriate number of degrees of freedom even when there is a real ranking effect. This is an example of a ' randomized decision procedure.'

Further discussion of these transformations or of other analyses of variance tests for ranked data appears to depend on the formulation of population models in which there is a real ranking effect. The lack of such models is, of course, a major obstacle to advances in almost all aspects of ranking theory.

REFERENCES

1. Fisher, R. A., and Yates, F. (1938). *Statistical Tables for Biological, Agricultural, and Medical Research.* London : Oliver & Boyd.
2. Quenouille, M. H. (1950). *Introductory Statistics.* London : Butterworth-Springer.
3. Bliss, C. I., Anderson, A. O., and Marland, R. E. (1943). ' A technique for testing consumer preference, with special reference to the constituents of ice-cream,' *Storrs Agric. Exp. Stat. Bulletin* 251, *Univ. of Connecticut.*
4. Pitman, E. J. G. (1938). ' Significance tests which may be applied to samples from any population. III,' *Biometrika*, XXX, 322–35.
5. Kendall, M. G., and Babington Smith, B. (1939). ' The problem of m rankings,' *Ann. Math. Stat.*, X, 275–87.
6. Kendall, M. G. (1948). *Rank Correlation Methods.* London : Griffin.
7. Friedman, M. (1940). ' A comparison of alternative tests of significance for the problem of m rankings,' *Ann. Math. Stat.*, XI, 86–92.
8. Mahalanobis, P. C., Bose, S. S., Roy, P. R., and Banerjee, S. K. (1933). ' Tables of random samples from a normal population,' *Sankhya*, I, 289–328.
9. Wold, H. (1948). *Random Normal Deviates.* Tracts for Computers, No. 25. Cambridge : University Press.

BOOK REVIEWS

Studies in Reading, Vol. II. Publications of the Scottish Council for Research in Education. No. XXXIV. London : University of London Press. 1950. Pp. 107. 7s. 6d. net.

The second volume of *Studies in Reading* contains the results of three investigations : the first deals with the relation between socio-economic status and reading ability ; the second with the effect of training on ' reading readiness ' ; and the third with the individual interests revealed by children's leisure-time reading. Of the three the first is the most ambitious. Its dual object was to " determine the effect of the socio-economic status of the home on reading attainment at the beginning stages, . . . and to compare the two most popular methods of teaching reading in the infant division of the Glasgow Primary Schools." The sample consisted of 1,080 children between the ages of 5 and 8, drawn from 45 schools. They were classified in eight groups, according to ' socio-economic level.' With one minor exception equal numbers of boys and girls at each age-level were drawn from each group ; and equal numbers were taught to read by one of three methods. Two were combinations of the Sentence and Phonic methods ; in the former, phonic training was given outside the regular reading period ; in the latter it was given as part of the reading lesson, but involved less emphasis on preparation for reading. The third ' method ' is said to cover all other procedures.

Specially devised tests of word recognition and reading comprehension were used. The results indicate that the method of teaching makes little difference. The investigator points out a few minor differences which emerge when age-groups are considered by themselves, but stresses the danger of drawing further conclusions from the data presented here. Although there was considerable overlap in attainment between the various socio-economic groups, there were significant differences between the mean scores of certain of the groups, the " percentage of attainment " decreasing with increase in the " percentage of necessity." The difficulties of interpreting these differences are rightly emphasized.

The Scottish Council is certainly to be congratulated on taking up so important a problem. But the research itself, though suggestive, is open to some criticism. As is so often the case, the references cited are confined almost entirely to American research, the results of which are only partly relevant to conditions in this country. The work of British psychologists is rather cursorily dismissed. Much space is devoted to American socio-economic scales ; but we are left in ignorance of the precise nature of the Glasgow scale on which the present enquiry is based. The teacher's interest will be aroused by the comparison of results attained by different methods of teaching. Yet we are told singularly little about these methods ; and the teacher searching for practical help may well feel that this study affords very little. The data were rather roughly analysed by estimating the critical ratio between the means. The investigation, however, would seem to be of the type particularly suitable for treatment by the method of analysis of variance—the procedure adopted, for instance, by Burt and Lewis in their research on ' Teaching Backward Readers ' (*Brit. J. Educ. Psychol.*, XVI, pp. 116–32).

The second paper deals with a more modest research carried out in four schools in Fife and two in Edinburgh. The main purpose was to investigate what difference an extra year of schooling at an early age has upon skill in reading. One hundred and fourteen children, who entered school at 5 years, were tested one year later with the Metropolitan Reading Readiness tests and the Metropolitan Achievement tests. Their mean scores on both were compared with those of American children of the same age, who would then be just beginning school : in the Reading Readiness test that of the Scottish children was higher by 16·3 points than that of the American ; in the Reading Achievement test the Scottish children were a year in advance. These results seem to indicate that children aged 5 years can make satisfactory progress in learning to read. Whether this is achieved at the cost of other values in education we do not know.

The third investigation shows that Glasgow children do not differ in their reading interests in any significant way from children in other towns in Britain or from children in the United States of America. It makes a distinct but somewhat limited contribution to the neglected field of British children's leisure-time interests, and may be heartily welcomed on that ground. *G. K.*

A List of Researches in Education and Educational Psychology. By A. M. Blackwell. National Foundation for Educational Research in England and Wales. Publication No. 1. London : Newnes Educational Publishing Co., Ltd. 1951. Pp. 173. £1 1s.

Both psychologists and educationists will be deeply indebted to Mrs. Blackwell for the compilation of this volume. It covers some 2,000 titles of theses accepted for the B.Ed., M.Ed., M.A., M.Sc., Ph.D., and other research degrees between 1918 and 1948, and deposited for the most part in University libraries. Educational psychologists have already found her lists, published in the *British Journal of Educational Psychology* (1943–5), invaluable for tracing previous work. The present volume not only adds more recent titles, but also covers the field of education in general ; so that it is about twice as long as the earlier publication. While occasionally minor errors spring to the eye, it is obvious that great care has been taken to secure accuracy and comprehensiveness. When theses have been published in full or summary form, references are usually given. But these are very far from complete. Possibly in preparing a future edition it might be worth while to make enquiries of University professors and lecturers, who should know if and where their past students' work has been published.

The great problem in such a compilation is how to classify and index. Mrs. Blackwell has adopted a modification of the Dewey decimal system, which furnishes a reasonably logical classification of subject matter. Within each division, titles are listed in order of date. Many are given under two or more divisions ; and an alphabetical index of subjects is provided. But there are still many difficulties. For example, the reviewer wished to trace two theses on personality assessments in school record cards, and only after prolonged search found them under Selection for Secondary Education. The section on Statistical Method lists scarcely a tithe of the theses containing interesting contributions to, or applications of, such methods. This is somewhat unfortunate : in the early days of work on educational psychology there were no facilities for publishing papers on statistical techniques or statistical enquiries, and hence many writers both in this country and abroad still overlook several of the pioneer investigations. Moreover, partly perhaps because of the ambiguity of titles, a number of theses are included which deal with general or industrial psychology or with pediatrics, and have no educational bearing. It would surely be helpful to have an author as well as a subject index. Less mental effort is often needed in recalling the name of a student than in pondering under what subject he is likely to have been classified. Both indices should give page rather than decimal references ; it is so much easier to look up p. 49 than 159·9285(3)(*a*).

It is noteworthy that the publication has been undertaken by the National Foundation for Educational Research in England and Wales, which will also be responsible for issuing periodic supplements. This gives a fine start to a series, which, it may be hoped, will some day rival the publications of the Scottish Council for Research in Education. *P. E. V.*

Vol. IV
Part III

The British Journal of Psychology,
Statistical Section

November,
1951

A FACTORIAL STUDY OF MATHEMATICAL ABILITIES

By M. K. BARAKAT
Institute of Education, Cairo

I. *Problem.* II. *Methods: A. Experimental; B. Statistical.* III. *Results: A. Influence of Cognitive Abilities; B. The Influence of Character Qualities; C. The Influence of Environment.* IV. *Summary and Conclusions.*

I. PROBLEM

Aim. The primary object of the investigation described in the following pages was to determine the nature of the chief cognitive factors affecting the progress of grammar school pupils in elementary mathematics. Supplementary enquiries were also included to discover what influence, if any, is attributable to affective or conative traits and to environmental circumstances such as conditions at home or in school.

Alternative Views. The earliest evidence for a group factor underlying work in arithmetic and mathematics appears to be that contained in Burt's investigations of children in London schools. He found that, among the various tests of elementary school subjects, problem work in arithmetic furnished one of the highest saturations for general ability or 'intelligence'; but, when the influence of this general factor had been partialled out, large and significant residuals still remained, so that the arithmetical tests revealed the "largest specific correlations of any of the groups of tests employed." [1] He insists that the 'factor' is by no means simple, and "probably involves a variety of more specific sub-factors: one appears to consist largely in an ability to remember and work with numerical symbols (as in mechanical arithmetic, like simple tests of the four rules); another in an ability to perceive and work with numerical relations (as in problem arithmetic)."

The existence of any such special aptitude was strongly denied by Spearman. As early as 1904, in discussing Wissler's results and his own, he declared that ability in mathematical subjects was determined by the same general ability as determined progress in all other subjects of the curriculum, and argued that his own data (marks in school examinations) "definitely disproved the opposite and more usual view, namely, that mathematics form an entirely independent faculty." [2] In a more recent publication, he again states that "contrary to popular belief (e.g., that of most mathematical teachers) there is *no* general mathematical ability: on eliminating *g* (the general factor), arithmetical and geometrical processes are found to show zero inter-correlations." [3] Thorndike and his followers deny that there are *any* general factors, whether broad or narrow: skill in arithmetic consists simply in "forming and strengthening an immense number of specific bonds or connexions between the sequence of numbers in counting, the items of the addition and multiplication tables, the rules of

[1] *Distribution and Relations of Educational Abilities,* 1917, pp. 52, and earlier L.C.C. Reports; also (7), (16), and (20).

[2] 'General Intelligence Objectively Measured and Determined,' *Amer. J. Psychol.*, XV, 1904, p. 276.

[3] *Psychology Down the Ages*, 1937, Vol. II, p. 253. In *The Abilities of Man*, after discussing the work of Collar and others carried out under Burt in the L.C.C. schools, he eventually agrees that there is some small evidence for a special arithmetical factor as an 'exceptional case'; but holds (with Thomson) that it is probably the result of differential experience or training. What *seems* to be a special ability for arithmetic may, he thinks, "perhaps be merely due to the extremely narrow basis (simple addition, subtraction, and multiplication) on which all arithmetical operations are ultimately founded"; in short, that it merely depends on whether or not the pupils have adequately acquired the necessary 'habit-memories' for the fundamental arithmetical tables. This is Thorndike's theory with a general factor added. Cf. also (8), p. 242.

algebra, the experiences of line and figure, and so forth." [1] In other words, as Stone expresses it, " there are as many different abilities as there are types of examples." [2]

Thurstone inclines to the view expressed by Burt. He agrees with Spearman that, on *a priori* grounds, we might expect differences in arithmetical achievements to belong to a " cultural rather than a biological category." But his own results, he says, lead him " with some confidence " to accept a special " Number factor." This he considers to be identical with the factor that has " appeared in many previous studies," and adds that it " almost certainly represents a unique ability." He differs from Burt in regarding it as a ' unitary ' or ' primary ability ' ; and observes that there " seems to be some evidence for the genetic interpretation of number facility as an inherited trait " (15, pp. 82–3). Thomson, on the other hand, would appear to favour the ' cultural ' rather than the ' biological ' interpretation. " Far from being divided up into unitary factors, the mind is a rich, comparatively undifferentiated complex of innumerable influences " : (cf. 25 and refs.).

These contradictory statements seem due partly to a misunderstanding of what a factor really is, partly to the inadequate planning of the investigations on which the conclusions are based, but most of all to a failure to combine statistical with other psychological evidence.

II. METHODS

A. Experimental. The Selection of Tests. Before we can plan a satisfactory experimental procedure, the problem to be investigated must be formulated in more specific terms. What abilities and traits are likely to be relevant ? And what are we to understand by ' mathematical ' ?

I. *Attainments.* For the purpose of selecting or devising the tests of attainments, it will be sufficient simply to say that by the term ' mathematics ' is meant those particular subjects which are generally taught in schools under that title. In discussing the curricula for grammar schools, the Board of Education's *Report* specifies three main branches—arithmetic, algebra, and geometry.

Accordingly, after studying current syllabuses, four tests were constructed and standardized, suited for pupils of the age and type of school here envisaged, namely, tests of (1) Arithmetical Computation ; (2) Arithmetical Problems ; (3) Algebra ; and (4) Geometry.[3]

II. *Abilities.* To decide what abilities are relevant to our problem is more difficult. In the past, narrow or mistaken preconceptions as to what mathematics really involves have led to misleading conclusions about the nature of the underlying factors. In selecting tests for the present research one general principle was found particularly helpful : " Before planning an experimental or statistical study of any concrete activity, it is wise to begin by a *theoretical* analysis of the psychological nature of the processes apparently involved, and to supplement this by a rough *empirical* study of the more salient differences between those who succeed and those who fail " (17). To start straight away with half-a-dozen or so convenient tests, chosen because they can be readily applied or represent popular ideas on the subject (as most previous investigators appear to have done), and then to seek a ' verbal,' ' spatial,' or ' numerical ' factor by suitable rotation—such a procedure can demonstrate little else besides the investigator's own predilections.

(a) *Theoretical Analysis.* Most psychological text-books include long discussions about the mental processes involved in verbal activities ; but nearly all are silent about the processes

[1] Cf. (5), p. 190 ; also *id.*, ' The Psychology of Drill in Arithmetic,' *J. Educ. Psychol.*, XIII, 1921, pp. 183–94.

[2] Stone, C. W., ' Arithmetical Abilities and the Factors determining them.' *Columbia Contributions to Education*, No. 19, 1908. P. Sandiford and F. E. Bolton accept the same view in their well-known text-books on educational psychology.

[3] Cf. (15), pp. 239–40. Copies of the test-sheets used will be found in the appendix to (24). To ensure that the problems included should be genuinely mathematical, as a mathematician would interpret the word, some consideration was also given to the conclusions reached in the following section : those conclusions seem closely in accord with the views expressed by Sir Percy Nunn (15, p. 451 and refs.), and quoted with approval in the Board's *Report* (*ibid.*, pp. 235f.). The deciding principle, however, was the teaching-syllabuses at the schools.

involved in numerical work. Yet it seems impossible to construct any valid tests for the abilities or factors involved unless we begin with some plausible hypothesis as to their nature.

The following analysis is taken mainly from Burt's review (12). He points out that, for any detailed information on the subject, we have to rely chiefly on those who have made a special study of the mental development of young children, especially writers like Sully, Valentine, and others, who have observed at first-hand how children acquire their notions of number during the pre-school stage, or those who, like Nunn and Burt himself, have attempted to find out how they actually attack mathematical problems in the classroom.[1] Factorists, in planning their experiments or in discussing their results, hardly ever refer to these psychological enquiries, and generally express their views in terms of the traditional doctrines current among mathematical writers in the nineteenth century.[2] The resulting conclusions naturally satisfy neither the modern mathematician nor the psychologist.

It will be helpful to consider first of all where the traditional account was at fault.[3] According to that doctrine, while other branches of knowledge are concerned with *qualities*, mathematics, pure or applied, is concerned with *quantities* (10). Thus, mathematics has commonly been defined as ' the science of quantity or magnitude.' As such it was considered to have two main branches—one concerned with *discrete* quantity and abstract or *numerical* magnitude (arithmetic, algebra, and their derivatives), and the other with *continuous* quantity and concrete or *extensional* magnitude (geometry and its derivatives) ; the former begins (so it was supposed) with temporal relations, as in counting, the latter with spatial relations, as in the estimation of length and shape (cf. 2, 10). Views like these have been accepted by most of the associationist writers who touch upon the problem, from Hume and Mill to Spearman and Thorndike : for detailed references, see Burt (12)).

Spearman, for example, seeks to explain these processes by postulating three relations which he calls ' space,' ' time ' (or sequence), and ' conjunction ' : (these are virtually Hume's ' relations ' of ' space,' ' time,' and ' number,' except that ' number ' is interpreted in terms of counting as the result of repeatedly ' conjoining ' or adding one).[4] Counting is a temporal process ; and so, we are told, quickly becomes mechanical as a result of temporal associations. Spatial concepts, being of two or more dimensions, commonly represent a more difficult and later development than number,[5] which is uni-dimensional ; hence geometry is held to be " more fully saturated with *g*." But for none of these is any new factor required : " all is effected by educing relations and correlates."

These and similar views have held the stage until quite recently. But both modern psychology and the modern theory of mathematics compel us urgently to re-examine the traditional arguments. In this country the development of the newer doctrines has been due chiefly to Russell and White-

[1] See Sully, J., ' Extracts from a Father's Diary,' in *Studies of Childhood*, 1896, pp. 468f. ; Bühler, K., *Die geistige Entwicklung des Kindes*, 1920, pp. 281f. ; Stern, W., *The Psychology of Early Childhood*, 1924, pp. 379f. ; Koffka, K., *The Growth of the Mind*, 1924, pp. 332–5, and, clearest of all, Valentine, C. W., *The Psychology of Early Childhood*, 1942, pp. 499f. Max Wertheimer's investigations of the arithmetical processes of primitive folk bring out much the same features : (' Uber des Denken der Natürvölker : I, Zahlen und Zahlgebilde,' *Zeitschr. f. Psychol.*, LX, 1912, pp. 320f.). For Nunn's views see his *Teaching of Algebra* (1914) and his chapter in *The New Teaching* (ed. Adams, 1918).

[2] Cf. Calkins, M. W., ' A Study of the Mathematical Consciousness,' *Educ. Rev.*, VIII, pp. 295f.

[3] In what follows I am much indebted to Professor Burt's *Notes on the Psychological Analysis of Mathematical Ability* (12), appended to one of his L.C.C. *Reports on Selection for Secondary Schools*, and thank both him and the London County Council for permission to quote from it.

[4] *The Nature of Intelligence and the Principles of Cognition* (1927), p. 72. Spearman's approach is still definitely associationist : his ' relations ' are simply conscious associations, and his classification of them is very similar to that given by Thorndike and other associationists (cf., for instance, Thorndike, E. L., *Elements of Psychology*, 1905, ' Classification of Relations,' pp. 61–2, and section on ' Logical Thought,' pp. 62f. : even ' eduction of correlates ' is simply a neater phrase for what Thorndike had called ' Response by Analogy '). Spearman, however, omits Thorndike's ' relation of equality ' and prefers to substitute the ' relation of conjunction.' This term he nowhere defines. Apparently he uses it to translate the German word *Hinzusetzung*, and is in effect adopting the general view of arithmetic and cardinal number, which was prevalent in Germany before the influence of Frege and Cantor had begun to spread : (cf. Meumann, E., *Experimentelle Padagogik*, 1907, II, chap. xvi, ' Das Rechnen,' or Rusk, R., *Experimental Education*, 1919, chap. xviii and refs. : and for criticisms see Young (10), pp. 64f., and Russell (2), pp. 11f.).

[5] This statement, which seems to have been made on purely *a priori* grounds, does not appear to be confirmed by experimental results : see, for example, Burt, C., ' The Development of Reasoning,' *J. Exp. Ped.*, V, 1919, pp. 125–7.

head. Their psychological implications were first taken up and applied to teaching and testing in secondary schools about 25 years ago by Nunn, Burt, and Hamley, who were the three Professors of Education at the London Day Training College (the Institute of Education in the University of London) at that time.

Let us consider the mathematical arguments first. What after all is number ? The popular contrast between quantity and quality obscures the fact that number, like red or sweet, is also an attribute. But it is an attribute that is predicable exclusively of collections, groups, or ' classes.' It has been defined as " that characteristic of a class or group which is common to the group itself and to all other groups whose individual members can be brought into a one-to-one correspondence with its individual members, and that regardless either of the nature of the individuals or of their arrangement as members of the group." [1] This definition, it will be observed, treats the fundamental relation, not as ' conjunction,' but as ' correspondence.'

This seems a definite improvement on Spearman's view : for, both logically and psychologically, the notion of similar or correspondent classes is prior to the notion of addition or counting. The ideas of ' variable ' and ' function,' which are basic in algebra, are also derived from the same two concepts—those of ' class ' and of ' correspondence ' respectively.[2] Or consider the modern treatment of geometry. Euclidean geometry, which Spearman has mainly in mind, is now regarded as merely a special application of projective geometry ; and the generalized ' points ' and ' lines ' of projective geometry have little or nothing to do with ' relations of space ' (in Spearman's sense of the word). Russell points out that they are more closely related to the idea of ' order.' But for Spearman, as for his immediate predecessors, order, like counting, arises from the perception of relations of time (or ' sequence,' as he sometimes calls it).[3] More recent observations and analysis, however, have made it perfectly clear that neither ' order ' nor ' sequence ' are themselves primitive notions ; they depend on the ability to recognize (implicitly at least) whether or not a generating relation is *asymmetrical, transitive, and connected* (13, pp. 521f., 535 ; 17, pp. 118f. and refs.).

This suggests, as Burt observes, that in their discussion of the psychology of mathematics the associationists (including both Spearman and his critics) are at fault in at least three respects.

(i) First, their conception of the *origin* of mathematical thinking is far too analytic and relational.

Spearman has always opposed " attempts at an interpretation in terms of synthesis or Gestalt " ; " here as elsewhere," he says, " the concept has only wrought confusion." [4] Yet both the philosophy of mathematics and the study of the pre-school child tend strongly to support the synthetic and schematic interpretations to which the Gestalt school and their forerunners inclined. In dealing with a pattern of five dots, for instance (as in Burt's spot pattern test), the associationists suppose that the child " makes five successive acts of attention," associating each dot with the last by a relation of sequence or conjunction. But, as experiment quickly shows, what is recognized by the child is primarily a *Gestalt-qualität*, which can (with tachistoscopic exposure) be distinguished with a single act of attention. According to Burt, therefore, the eventual discovery of number depends in the first instance on grouping rather than on serialization ; and, in accepting this synthetic interpretation, he makes

[1] The substance of the definition is due to the revolutionary notions put forward in G. Cantor's *Mengenlehre* (Theory of Classes, 1872) : the first school text-book to adopt it was Fine's *College Algebra* (Princeton, 1905). The wording in the text is Burt's (12).

[2] Cf. (3) and (10). Piaget in his recent volume now adopts much the same view. " The operations constituting number," he says, depend on " one-to-one correspondence, with conservation of the resulting equivalence, despite differences in pattern or shape." However, he seems to think that counting and " ordering " (" simple repetition of unity, $1 + 1 = 2$, $2 + 1 = 3$, etc.") develop at much the same stage as the intuitive discrimination of small numbered groups (e.g., couples, trios, or quadruplets) : see *The Psychology of Intelligence*, 1950, pp. 144f.

[3] The doctrine that " time, as a ' form of intuition,' is the basis of number," is also accepted by James (*Principles*, II, p. 653 : " Number seems primarily to signify the strokes of our attention in counting things "). It is instructive to find that this view has been criticized in almost precisely the same way both by mathematicians (e.g., Young, 10, p. 64) and by psychologists (e.g., Ward, *Psychological Principles*, 1918, pp. 323–4 : " counting does not make number ; if it is to be described as ' consisting in the colligation of separate acts of attention ' it does not even ascertain number . . . It is in no way essential to it." And again " number does not arise from counting ; . . . simultaneous (spatial) plurality is a better provocative than successive (temporal) ").

[4] *The Nature of Intelligence*, p. 261 ; *The Abilities of Man*, p. 21.

apperception rather than association or relating-educing the earlier and indeed the central process.[1]

Counting is a comparatively late and complex achievement. Spearman, Rusk, and others who treat counting as basic greatly over-simplify the psychological processes involved. In counting the child has to establish a factitious ' correspondence ' between the individual items in a group and a fixed set of number names temporarily attached to those items. This implies that he has previously memorized, in a temporal order by sheer mechanical association, a series of meaningless proper names (e.g., in Greek the letters of the alphabet ; in later languages the names of the first 12 Arabic symbols). To effect an accurate correspondence, the enumeration has to be both exhaustive and non-repetitive.[2] Only when the child can correctly answer the question ' How many ? ' (at least for tiny groups) can it be said to have acquired a clear conception of number. Thus, as Ward puts it, the concept of ' quotity ' is even more essential than that of ' quantity.'

(ii) Secondly, the associationists retain erroneous and out-of-date notions of the typical *operations* of mathematics. What distinguishes mathematics from other sciences is its method of proof : this is essentially formal. Hence modern views describe mathematics, not as a science or group of sciences distinguished by a special subject-matter (geometry, for example, being the science of space), but as a branch of logic,[3] concerned primarily with formal processes.

It is this that makes it so different from all the other topics of the school curriculum, and introduces special difficulties for so many pupils.[4] Certainly, Spearman's noegenetic treatment here marks a definite advance on that of the older associationists ; but his own tests of mathematical reasoning show how incomplete that theory is for the purpose intended. From a psychological standpoint, as Burt shows, " tests of logical reasoning involve three fundamental cognitive processes, and have to answer three questions. Can the child *analyse* a concrete whole and educe the implicit relations ? Can he *apply* the relations and so educe correlates ? And can he *recombine* these relations and correlates in a complex scheme, organized in accordance with the appropriate ' form of synthesis ' " ? [5] Now, in formulating his ' noegenetic laws,' Spearman appears only to have taken Burt's first two processes or abilities—the ability to perceive isolated relations and the ability to apply isolated relations : the importance of the third—the use of the appropriate logical scheme or ' form of synthesis ' (Stout's phrase), a process which is essential to any concrete logical operation—he almost wholly ignores. Both for logic and for psychology, " the important thing about relations is not their material but their formal classification " (12) : the difficulty for the child is not so much (as Spearman supposes) to distinguish a relation of ' conjunction ' from one of ' constitution,' or of ' space ' from one of ' time,' but to " distinguish asymmetrical relations from symmetrical, transitive from intransitive, connected from unconnected, many-one from one-one," and so on. This is where the pupil who fails in mathematics so often breaks down.

(iii) Thirdly, the associationists' notion of the *content* of mathematics is at once too simple and too concrete. It is supposed to deal specifically with relations of space, time, and sequence ; and these were classified by Spearman and earlier associationists as ' real ' or concrete relations, in contrast with the ' ideal ' or abstract. According to this view all

[1] The point is of some practical importance. Both in Great Britain and in Germany there has for long been considerable controversy between two different approaches to the teaching of arithmetic. The academic psychologist has, until recently, favoured what, in Germany pedagogy, is called the *Zählmethode*, i.e., the analytic approach ; the practical teacher more commonly favours the *Anschaungsbildermethode*, i.e., the synthetic approach. (Cf. Meumann, E., *loc. cit. sup.*, pp. 331f. ; also the experiments of H. A. Nanu, *Zur Psychologie der Zahlauffassung*, 1904. Freeman, F. N., ' Untersuchungen über den Aufmerksamkeitsumfang und die Zahlauffassung,' *Päd. Psychol. Arbeit.*, I, 1906, pp. 88–168 : cf. also (11) and (13).

[2] Everyone who has used Binet's test of counting knows how difficult for the tiny child these two conditions are, and how hard it is for him to realize that the correspondence between things and names are temporary, arbitrary, in short, ' factitious not factual.'

[3] But only a branch. Peirce's famous definition of mathematics as ' the science which draws necessary conclusions ' is (as Young and others have observed) at once too broad (it states only the genus) and too narrow (' formal logic is not the only method in mathematics ') : cf. 10, pp. 217, 220. Also Bôcher, M., ' The Fundamental Methods of Mathematics,' *Bull. Amer. Math. Soc.*, XI, pp. 115f.

[4] See (11) and the experiments by Miss Johnston and Branford described in (16), pp. 444, 449.

[5] Burt, C., ' The Development of Reasoning in Children,' *J. Exp. Ped.*, V, 1919, pp. 121–7, and (13), p. 521. The two principles of perceiving or ' educing ' a relation and applying it to obtain a correlate (as exemplified in his Analogies tests) were described by Burt in *J. Exp. Ped.*, I, 1911, p. 101 ; but he supplements this by his ' Syllogistic test,' which completes the synthesis (pp. 102, 104). In mathematics the typical syllogism is, of course, the ' functional syllogism ' : see Johnson, W. E., *Logic*, Pt. II, ch. V.

the child has to do is to apply the ' noegenetic principles of intelligence ' to a content embodying these relations. Burt, on the other hand, maintains that the content is not ' real ' or concrete, but is itself formal : " the distinctive peculiarity of mathematics, as contrasted with other branches of logic, is that it introduces, not only material *variables* and *formal* constants, but also *formal variables*."

From a psychological standpoint it would seem impossible to explain how we can apprehend ' *formal* variables ' unless we accept something like Stout's doctrine of apperceptive schemes or the Gestalt school's doctrine of forms or configurations.[1] The whole mental process is greatly facilitated by the use of symbols ; and a striking omission from the tests used by Thorndike, Spearman, and their disciples are tests which will show how readily the child can work with formal schemes, and grasp, use, and manipulate substitute-symbols.[2] The factorial study, recently published by Coombs, appears largely to confirm this final conclusion. But he seems to go too far when he ends by declaring that " [his] results are in agreement with the hypothesis that number ability is characterized by *facility in manipulating a symbolic system according to a given rule*, and apparently offers this as a sufficient as well as a necessary description."[3]

The foregoing theoretical analysis makes it clear that the batteries of tests used by Thorndike, Spearman, Thurstone, and their disciples cover only a small part of the total field, and that not the most essential. A complete investigation of the components of mathematical ability would require tests for all the various mental processes mentioned in this preliminary review. But for a preliminary study it was deemed sufficient to take only those that appear relevant to progress in mathematical subjects at the secondary school stage. In attempting such a choice we may gain further help if we turn from this theoretical analysis to more empirical investigations.

(b) *Empirical Analysis.* In studying the causes of mental subnormality it has often been urged that statistical methods should be supplemented or preceded by case-study methods. The same holds good for the study of supernormal ability. In the field of mathematics the most systematic attempt to apply the case-study method is that undertaken by Burt (in conjunction with several research students) during their work in the London schools (11 : cf. also 9).

A comparative investigation was made of four main groups : (*a*) 114 pupils with specific disability in arithmetic (under 85 A.Q.) ; (*b*) 87 with high mathematical ability (over 115 A.Q.) ; (*c*) 114 pupils, of average ability in mathematical subjects, matched with the first for sex, age, and intelligence, and (*d*) 87 matched with the second. An analysis of variance and covariance was carried out ; and differences in the following characteristics were found to be statistically significant. They are expressed as they characterize the second group, and arranged roughly in order of significance. (i) Good long-distance mechanical memory ; (ii) High intelligence : (with 3 exceptions, all children showing high mathematical ability had I.Q.s over 125 ; the poor arithmeticians had I.Q.s between

[1] Cf. Burt's appendix to the Board of Education *Report* (15), pp. 435–6. The main difference between the two writers seems to be that Burt accepts, and Spearman rejects, Stout's well-known solution to the problem : " Can the form of a whole be apprehended without apprehension of its concrete parts ? " Spearman *a fortiori* has to reject Stout's further conclusion that " in general the presentation of the Form of Synthesis is *psychologically* more primitive than that of the Relations " (*Analytic Psychology*, I, p. 73). Spearman indeed complains that, owing to insufficient analysis, Stout " has not discovered the dominant part played by the relations : the very word ' relation ' is hardly mentioned by him " (a surprising statement, since in Stout's section " On Relations " the word is mentioned more than thirty times).

[2] In test-construction it is important to note that psychologically symbols are of two main kinds. A *word* is an ' expressive symbol,' whereas a *number* (or rather a numeral) is a ' substitute symbol ' : it is " a sign or counter which takes the place of the idea named." " A word is a means of *thinking about* its meaning ; a substitute symbol is a means of *not* thinking about the meaning " : (Stout, G. F., *Analytic Psychology*, Vol. II, pp. 193–4).

[3] Coombs, C. H., ' A Factorial Study of Number Ability,' *Psychometrika*, VI (1941), p. 189. Interesting as they are, Coombs' final conclusions do not appear to follow from his factorial analysis. He himself observes : " The rotated factorial matrix does not yield any information on the hypotheses the batteries were designed to investigate, because the primaries are oblique." Consequently, he falls back on references to certain of the correlation coefficients as originally observed. But those who believe in the existence of a general factor will feel that an unknown proportion of these correlations may be due, not to the influence of number ability as such, but to ' general intelligence.'

73 and 117) ; (iii) Wide span of attention[1] ; (iv) Emotional stability ; (v) Introverted[2] rather than extraverted tendencies ; (vi) Ease in apperceiving formal schemes, patterns, and designs (e.g., in spot pattern tests, matrix tests, and tests of perceptual speed) ; (vii) A marked interest in numbers and their properties, displayed at a very early age :[3] (the poor arithmeticians were, as a rule, not merely bored by numbers, but evinced an emotional antipathy to figures) ; (viii) Ability to reason deductively, particularly where functional relations rather than subsumptive relations are involved : (this means drawing formal conclusions from formal premises as distinct from drawing factual conclusions from factual premises) ; (ix) Ability to reason inductively with formal material (e.g., ability to detect and apply the formula underlying a progression or a sequence) ; (x) Ability to detect and apply implicit relations (as in verbal and non-verbal analogies) ; (xi) Audito-motor imagery (the majority of those with high mathematical ability were of a verbal rather than a concrete or pictorial type ; but none had disproportionately high assessments for verbal facility, whereas the backward arithmeticians included many verbalizers ; marked lack of visual imagery appeared a handicap in geometry and occasionally in problems in applied arithmetic) ; (xii) Ease in using substitute symbols according to arbitrary schemes (e.g., in a code test) ; (xiii) And generally, a readiness for an abstract, formal, symbolic mode of thinking rather than a concrete, material, or linguistic mode of thinking : (it is emphasized that the process is often implicit and intuitive). Thus, " so far from being itself a simple or primary capacity, and quite apart from any acquired skill in dealing with numerical content, what is commonly called ' mathematical ability ' appears largely to depend on a complex combination or conjunction of several comparatively elementary cognitive processes."

Burt further notes that " there seem to be two main sub-types : among the good mathematicians, both in the secondary school and in the university, it is possible to distinguish (*a*) the analytic or explicit thinkers, who work by slow, logical reasoning (they are usually introverts and nearly always of an audito-motor and verbal type, and take easily to algebra and analysis), and (*b*) the synthetic or intuitive thinkers, who work by quick intuitive methods (they are often poor in verbal imagery, and not infrequently are visualizers and motiles rather than audiles ; as a rule they are better at geometry and trigonometry than arithmetic and algebra, and, even for non-geometrical problems, prefer geometrical formulations and proofs)." [4] Only the first of these subtypes are envisaged in tests based on Spearman's noegenetic principles.

Tests of Ability and Aptitude. The foregoing considerations primarily determined the choice of mental tests for the present research. In pilot surveys a large number were tried out both individually and in a grammar school. The final series, as used in the main experiment, was limited to the following :

(1) Intelligence (*a*) verbal and (*b*) non-verbal (viz., Cattell's Test of Verbal Analogies and the N.I.I.P. Group Test No. 70/23). This covered point ii in Burt's list. (2) Verbal Reasoning (viz., Test VII from Burt's Northumberland Test). This covered viii, ix, and x. (3) Mechanical Memory (*a*) for Nonsense Syllables, (*b*) for Numbers (both intended to test span as well as memory), (*c*) for Spatial Patterns (N.I.I.P. test). These covered i, iii, and vi. (4) Letter Sequences (intended largely as a test of inductive capacity, i.e., discovering a rule or formula). This covered vi, ix; and, to some extent, xii and xiii. (5) Number Addition. This and 2*b* covered vii. Test 5 and the last were modified from Thurstone's tests for ' Primary Mental Abilities ' (14), pp. 44-6. (6) Visuo-Spatial

[1] This point is noted by Binet. He found that, in one set of experiments, Inaudi was able to repeat 36 figures after a single hearing (*Psychologie des grands calculateurs*, 1894, pp. 24f.). Burt, however, points out that, with his young mathematicians, the span for figures greatly exceeded the span for letters, and explains that with them many combinations of figures come to have a meaning almost like words—they are as familiar as squares, cubes, reciprocals, etc.

[2] The invention and solution of numerical problems is a favourite pastime with young introverts and lonely boys. For other activities the child needs companions, playgrounds, books, toys, paintboxes, or musical instruments. Of all forms of knowledge or skill, mathematics is thus the most independent and self-sufficing. Burt adds : " Among the London cases 11·5 per cent. of the mathematicians were only children, as compared with 2·3 per cent. in the other groups ; a number lived isolated lives." The few showing sthenic, assertive, or extravertive tendencies (often repressed) were inclined to obsessions or compulsions—particularly with numbers.

[3] We are told how the father of one precocious boy recalled that " when Jacob was five or thereabouts, he used to play with numbers as other children play with dolls or bricks : his discovery that the reciprocal of 81 was 12345 positively thrilled him."

[4] He refers to Henri Poincaré's well-known discussion of ' l'invention mathématique ' in the *Bulletin de l'Institut Général Psychologique* for 1908, and recalls the amusing controversies of a century ago ; for instance, the editors of *Liouville's Journal* would only admit algebraic proofs, even of geometrical theorems, whereas *Crelle's Journal* always insisted on geometrical proofs. He notes how easily the reader can discern these two opposite tendencies in different writers on statistics and factor analysis. (Observe, for example, how Thurstone complains of the lack of diagrams in Burt's book ; and, as Burt has told us, he himself is of audile rather than a visualizing type : *How the Mind Works*, p. 45.)

Perception (N.I.I.P. Group Test, No. 81). This partly covered vi and x, but was introduced mainly for its possible relation to geometrical ability and to type : (*b*) in Burt's analysis.[1]

Subjects. Until quite recently almost all the psychological studies of mathematical subjects have been carried out on pupils of the elementary or primary schools ; very few have been attempted at the grammar school stage.[2] Partly for this reason, and partly because of the practical importance of selecting pupils for different types of secondary school, the present investigation was undertaken with pupils attending London grammar schools. After a number of preliminary studies a trial enquiry was carried out on 52 pupils ; and finally in the main investigation (which alone will be reported here) the pupils tested comprised 160 boys and 160 girls, aged from $13\frac{1}{2}$ to $14\frac{1}{2}$ years, attending grammar schools in four different districts.

B. Statistical. (a) *Factor Analysis.* If we possessed a standard battery to cover all cognitive abilities, no doubt the appropriate method would be to factorize the ability tests first of all, and then correlate the attainments tests with the factors so found. But since practical conditions have limited us to a tentative selection of a relatively small number of tests, the best way of verifying our hypothesis will be to incorporate the attainments tests in the matrix to be factorized.

The correlation tables were first analysed by Burt's ' simple summation ' formula (1, p. 53 ; 17, p. 461). Five factors were extracted, but only three were found to be significant.

The formula used is identical with what Thurstone has since called the ' centroid formula ' ; but the two procedures involve three minor differences. (i) In the diagonal of the correlation matrix Thurstone inserts the highest figure in each column (whether observed coefficient or residual). Burt inserts as trial values figures that are slightly larger than these in those columns that yield the highest totals and slightly smaller in those columns that yield the lowest totals. Thurstone's rule is easier in the sense that it is more mechanical. But it is apt to magnify the number and size of the later factors, and to upset the equality of the positive and negative sections.[3] When successive approximations are employed (in order to ensure that the resulting self-correlations shall agree with the assumed self-correlations) Burt's method in the end saves time. (ii) Thurstone's aim is to " maintain the rank of the side correlations as the minimum possible rank "—surely a very irrelevant way of determining the number of factors (19, p. 283). Burt's is to discover the minimum number of significant factors required to account for the observed coefficients. (iii) For the bipolar factors Thurstone bases his sign-changes on a separate tabulation—a rather needless and laborious process. Burt rearranges the table to bring positive figures (so far as possible) into the top left quadrant and bottom right, and negative into the top right and bottom left. This is in accordance with his view of simple summation as a convenient substitute for weighted summation. With data such as the present the modification termed the method of ' subdivided factors ' (this *Journal*, II, p. 54) gives a much more plausible result, and was adopted after the first two factors (general and first bipolar) had been extracted.

Relative Merits of Graphic and Arithmetical Rotation. Since the factors are conceived as positive influences, more or less overlapping with one another, the factorial matrix that we eventually may expect to find (if our guiding hypothesis is true) will be a matrix of significant overlapping group factors, with one factor perhaps overlapping all the others so completely as to become a ' general factor.'[4]

[1] Tests of mental imagery were also included in the pilot survey (to cover point vii), but were found to have a low reliability and to be too time-consuming ; the attempt to test long-distance memory and memory for meaningful material had to be abandoned for the same reasons.

[2] An exception should be made of several studies carried out under Professors Nunn and Burt at the London Day Training College (a college for the training of teachers in secondary schools). Unfortunately very few of these have been published ; but a fairly complete list is now available in the recently published *List of Researches in Education and Educational Psychology* (1951) by A. M. Blackwell. A critical summary of the more important investigations will be found in my thesis (24, pp. 181f.).

[3] An excellent instance of this has since been provided by Sutherland's centroid analysis in a recent issue of this *Journal* (IV, Table III, p. 23). Instead of the factors diminishing in size after the third, it will be noted that factor IV is actually larger than factor III.

[4] Recent critics of factor analysis have frequently objected that the choice between possible procedures is " arbitrary." " There is nothing," we are told, " to help us decide whether our so-called factors are ' bipolar,' ' group,' or of some other type " (at least when the numbers tested are small, a reservation often overlooked). But this, like so many criticisms of mathematical procedures

The speediest way of securing such a factor matrix is by rotation from a preliminary summational matrix. Two procedures are in common use—Burt's method of arithmetical rotation and Thurstone's method of graphic rotation. There appears to have been little or no systematic study [1] of the comparative merits of the two procedures. Accordingly, with the first set of data (that obtained from the boys) both methods were tried.

For the graphical procedure Thurstone's instructions[2] were carefully followed. The 'centroid' factors were rotated blindly, two at a time. With 5 summational factors 10 preliminary diagrams (cf. 15, p. 73) are necessary before deciding with which pair to begin: the only instruction which Thurstone gives for deciding between the various pairs is to "examine each plot to determine whether the rotation will (1) increase the number of nearly vanishing entries, or (2) decrease the number of negative projections or (3) their absolute magnitudes." With these three criteria to balance against each other, and no means of judging except mere inspection of the graphs, the final choice is bound to be highly speculative. The selection of the first pair was relatively easy; but thereafter several blind trials were needed at each stage, and the decision between the alternatives became increasingly arbitrary and subjective. On showing the plots to different colleagues it was found that, quite frequently, different selections were proposed; some laid greatest emphasis on reducing negative values, others on increasing the zero entries. And those who remembered the nature of the tests were obviously biased, often it would seem unwittingly, by their knowledge and their preconceptions.

Burt's procedure is objective and impersonal. He begins, as Thurstone does, with an analysis by simple summation. But, whereas Thurstone then ignores the resulting sign-pattern, Burt uses it to decide the lines of demarcation between the various groups. With the present set of data, as we shall see, the slight modification requisite to obtain subdivided factors brings out the essential grouping still more clearly. With the lines of demarcation thus determined, a *direct* group factor analysis is next carried out. This yields non-overlapping group factors. And from them a rotation matrix can rapidly be constructed by solving a small number of simple linear equations.[3]

After carefully trying out the two methods, it was concluded that the arithmetical procedure is unquestionably to be preferred to the graphical, and that on four main grounds.

(i) The most obvious difference between the two lies in the *superior ease and speed* of the arithmetical method, when more than one series of graphical rotations is likely to be required.[4]

(ii) But secondly, as we have seen, with the graphical method, the choice of processes at each successive stage is by no means unique. It follows that the results achieved will be by no means unique. Since correlated factors are allowed, three or four quite different hypotheses can easily be confirmed. Indeed, when studying the rotated factors obtained by Thurstone or his followers, and wondering

in psychology, would apply equally well, if valid, to similar procedures elsewhere. The choice is merely one instance of the familiar choice between different ways of partitioning the degrees of freedom. Take, for instance, the illustration given by Jeans to exemplify the alternatives available in statistical mechanics. "A diatomic molecule has six degrees of freedom. We might take the six to consist (i) of the three degrees of freedom of the centroid of the molecule to move in space" (the general factor), "(ii) of the one degree of freedom arising from the possibility that the two atoms might change their distance apart" (a bipolar factor), "and (iii) the two degrees of freedom which allow the line joining the two atoms to change its direction" (which could be regarded as two specific factors). "Alternatively, we may suppose each atom to have its own three degrees of freedom" (two independent group factors). In statistical psychology, as in statistical mechanics, the actual choice will be determined by external considerations, e.g., by the form of the question we set out to answer: (cf. Jeans, J., *Kinetic Theory of Gases*, p. 254: the example and the reference are taken from (12). For a simple discussion of the general principles underlying the partitioning of degrees of freedom, see Mather, K., *Statistical Analysis in Biology*, pp. 61–8).

[1] I have since learnt that, in an unpublished thesis, describing a factorial analysis of data from the Australian Air Force, Mr. D. K. Wheeler has also applied both methods, and found Thurstone's procedure somewhat unsatisfactory. In this *Journal* Raath and Reyburn have strongly criticized Thurstone's method of rotation (II, pp. 125f., III, pp. 150f.): they quote and endorse the criticisms made by Burt (*Factors of the Mind*, pp. 358f.); they themselves do not find the results of Thurstone's procedure by any means unique, and complain that, particularly when oblique factors are permitted, a subjective or arbitrary choice is very apt to intervene, consciously or unconsciously, at the various successive stages: on the other hand, the method they themselves propose seems still more cumbersome.

[2] See my thesis, pp. 126f.; also (19), pp. 319–409, and (15), pp. 73f.

[3] See Burt, C., 'Group Factor Analysis,' this *Journal*, III, Appendix on Working Methods, p. 75. A more elementary account will be found in Burt's 1929 memorandum (12) and in *Journ. Psychol.*, VI, 1938, pp. 339f.

[4] This conclusion seems further confirmed by the recent study of Thurstone's procedures published in the last issue of this *Journal* by J. Sutherland. The reader will readily agree when he asks "what is gained" by the elaborate succession of procedures that he describes, and concludes that "in most experiments it is hardly justifiable to go beyond the original orthogonal or oblique factors" (p. 29).

whether or not to accept that preferred by the investigator,[1] does not almost every reader turn back to the bipolar (or ' centroid ') analysis to see which grouping is really indicated ? Burt's procedure, on the other hand, is mechanical at every stage. Consequently, the outcome is objectively determined, and is in general *unique*. Every computer is bound to reach the same results.[2]

(iii) The group factors reached by the arithmetical procedure are, as a rule, *easy to interpret*. On the other hand, when a Thurstone rotation is carried out with nothing but the numerical data to guide us (i.e., with a due disregard of the psychological nature of the tests involved), then, as many critics have noted, the resulting factors frequently permit of only a dubious and confused interpretation. In the present research a glance at the figures resulting first from blind rotation (24, pp. 128f.) and then at those resulting from further rotations *with* knowledge of the tests, shows how difficult to interpret the former are. Consequently, in actual practice nearly every investigator who follows Thurstone's instructions does *not* blindly disregard the nature of the tests : he deliberately selects his axes so as to secure factors that will be (in Thurstone's phrase) " psychologically meaningful." [3] And, of course, these commonly turn out to be the factors with that particular ' psychological meaning ' which the investigator himself privately favours. But surely this is no way to verify a scientific hypothesis. To give the hypothesis a fair chance of being disproved, the analysis, rotation and all, should be entirely automatic from start to finish.

(iv) Finally, Burt claims that, with his procedure, the group factor analysis is likely to correct any artificiality or variability the bipolar factors may have as a result of an ill-balanced selection of tests or traits. As a consequence, the group factors thus obtained are not only objectively determined, but also *remain invariant* when other tests are added to the battery. So far as my own results go, all these claims seem fully justified.

Whichever procedure is adopted, it should be emphasized that, as a rule, a group factor-matrix is confirmatory only. Unless the experiment has been badly planned, the essential lines on which the tests are to be classified and sub-classified should emerge quite clearly in the preliminary bipolar analysis. Under these conditions the group factor analysis is required merely to show how far the resulting principles of classification can be safely interpreted as positive factors, and rarely adds much new information.

(b) *Analysis of Variance.* In investigating the influence of non-graded characteristics—such as differences in sex, school, occupational status of parents—an analysis of variance seemed the more appropriate procedure. In the main the working methods described by Lindquist (18) were followed : (for fuller details, see 24).

III. RESULTS

A. Cognitive Abilities. Inter-correlations were calculated between the thirteen tests for the two sexes separately (for detailed tables, see 24, p. 74). The final results of the factorial analyses are shown in Table I. The order of the tests has been rearranged to exhibit the classification revealed by the factorial results.

As will be seen, the classifications suggested by the figures for the two sexes are, on the whole, remarkably consistent. With both boys and girls the preliminary summational analysis yields three significant factors.

(i) The first is a general factor common to all the tests : it contributes rather less to the total variance than usual, because the pupils tested had already been selected for admission to a grammar

[1] Coombs, for example (*loc. cit. sup.*), with a sample of only 223 pupils obtains ten rotated factors ; in the end, however, he selects only seven. But which are to be selected ? A glance at his bipolar factors (Table 4), preferably when the tests have been rearranged to exhibit the factor-pattern, would surely leave the reader in far less doubt. Sutherland leaves the reader to choose between at least three alternative interpretations (' Summary,' p. 30) : how much easier the choice would have been had his Table II been rearranged to bring out the bipolar classification !

[2] Since the group factor-pattern which expresses a given bipolar pattern is of necessity a factor-matrix of higher rank than the latter, there may at first sight seem room for personal preferences. But, according to the strict procedure, all questions of this kind should be settled by tests of statistical significance. Thus, in the tables here factorized, we might have expected a broad mathematical group factor to appear (corresponding to the first bipolar, i.e., factor II) as well as the two narrower factors (corresponding to the second portion of the second bipolar, i.e., factor III). Owing to the small size of the sample, however, the broader factor proved to be non-significant, while leaving significant residuals. This, however, is a defect in the experimental design (choice of too small a sample), not in the statistical method. Some degree of doubt may also arise in factorizing correlation tables where the selection of tests or traits has not been carefully planned to verify or disprove a clear-cut psychological hypothesis.

[3] Sutherland, for example, says quite frankly that the positions given by his table of factors " are not the only possible positions : they are the positions which, in the opinion of the author, give the most satisfactory psychological interpretations " (p. 23).

TABLE I. FACTOR-SATURATIONS FOR MATHEMATICAL AND NON-MATHEMATICAL TESTS

Tests	Summational Factors					Group Factors									
	Undivided			Divided		Non-Overlapping					Rotated				
A. Boys	I	II	III	IIIA	IIIB	G	M	N	V	K	G	M	N	V	K
1. Algebra	·713	·118	·212	·176		·683	·237				·721	·147	—·032	·205	·017
2. Geometry	·647	·102	·026	·164		·745	·251				·625	·128	—·095	·006	·175
3. Arithmetic : Problems	·749	·119	·246	·258		·764	·256				·744	·237	—·124	·146	·022
4. Arithmetic : Mechanical	·676	·336	·127	·191		·686	·125	·213			·719	·246	·102	—·048	—·031
5. Numerical Addition	·498	·341	—·014	—·092		·441		·402			·487	·128	·317	·083	—·134
6. Memory : Numbers	·319	·175	—·174	—·370		·259		·563			·273	·114	·309	·010	·037
7. Memory : Nonsense Syllables	·373	·214	—·230	—·176		·332		·270			·357	·223	·210	—·123	·148
8. Letter Sequences	·576	·047	—·193	—·151		·584		·216			·595	·084	·186	·076	·025
9. Analogies : Verbal	·429	—·127	·345		·387	·411			·397		·352	·145	—·132	·393	·014
10. Reasoning : Verbal	·514	—·183	·208		·278	·489			·405		·403	·013	·054	·435	·049
11. Intelligence : Non-Verbal	·583	—·314	—·188		—·165	·514				·424	·513	—·098	·086	·168	·414
12. Visual Space	·629	—·543	—·141		—·326	·511				·684	·437	—·006	·013	·255	·674
13. Memory : Designs	·563	—·285	—·224		—·174	·486				·495	·441	·047	·008	·096	·497
Contribution to Variance (per cent.)	32·9	6·7	3·9	2·8	3·0	30·4	1·5	4·9	2·5	6·9	28·8	2·1	2·7	4·1	7·3
B. Girls	I	II	III	IIIA	IIIB	G	M	N	V	K	G	M	N	V	K
1. Algebra	·669	·243	·033	·103		·679	·145				·685	·187	·123	—·046	·112
2. Geometry	·704	·086	·269	·261		·718	·212				·678	·247	·008	—·094	·217
3. Arithmetic : Problems	·691	·173	·217	·258		·686	·528				·660	·381	·067	—·032	·029
4. Arithmetic : Mechanical	·637	·149	—·144	—·105		·629		·202			·553	·171	·284	·156	—·087
5. Numerical Addition	·419	·461	—·287	—·404		·343		·847			·316	·162	·601	·097	—·113
6. Memory : Numbers	·373	·098	—·088	—·113		·358		·193			·372	—·023	·217	—·114	·027
7. Memory : Nonsense Syllables	·355	—·137	—·202		—·026	·346			·062		·334	—·104	·163	·125	·148
8. Letter Sequences	·576	—·092	—·155		—·133	·575			·113		·549	·022	·057	·253	·054
9. Analogies : Verbal	·478	—·056	—·063		—·272	·437			·561		·451	·096	—·084	·329	—·052
10. Reasoning : Verbal	·432	—·061	—·137		—·216	·401			·217		·426	—·088	—·028	·186	—·045
11. Intelligence : Non-Verbal	·671	—·196	·116		·155	·642				·303	·594	·187	·012	·129	·328
12. Visual Space	·537	—·408	·285		·370	·439				·738	·425	·089	—·057	·031	·663
13. Memory : Designs	·675	—·260	·156		·122	·636				·301	·648	—·079	—·086	—·067	·333
Contribution to Variance (per cent.)	32·4	5·0	3·4	2·6	2·4	30·1	2·7	6·1	2·9	7·0	28·1	2·9	4·3	2·3	5·9

school. (ii) The second factor (the first bipolar) divides the tests into two main classes—roughly, as we shall see, the mathematical and the non-mathematical. (iii) The third factor subdivides each of these classes into two subclasses. Here, it will be observed, there are certain minor differences between the results for the two sexes: the two subgroups which get saturations of similar and negative signs with the girls get saturations of opposite signs with the boys. Now on referring to the table of residuals on which this third factor is based, it appears that the cross-correlations for these subgroups are for the most part non-significant. In such a case, it is misleading [1] to force one and the same factor to account for four distinct subclasses; and accordingly the method of subdivided factors was applied to this residual matrix. With both procedures, however, the main lines of classification are clear. With both the boys and the girls the non-mathematical tests subdivided into verbal and non-verbal (i.e., spatial), and the mathematical into mechanical or arithmetical and logical or mathematical (in the more specialized sense); but the lines of division are somewhat different with the two sexes.

Owing to the comparatively small numbers available, it was impossible to apply the method of subdivided factors in the group factor analysis: that would have meant extracting at least six independent factors from a correlation matrix whose significant rank could not be much more than three. The simplest procedure would be to undertake a group factor analysis, based on a threefold classification only—verbal, spatial, and mathematical. This gave a plausible fit; but, as already noted, it left small clusters of residuals in the mathematical tests, and the broad mathematical factor itself was not fully significant. Since the subdivision of the mathematical tests is a point that has been vigorously challenged, it seemed more advisable to study the effect of splitting the mathematical factor into two, as the second of the two bipolars suggests. This furnished a closer fit, fully significant factors, and no significant residuals.

Interpretation. As Burt has insisted, the interpretation of the factors found should not be tacked on as an afterthought to explain the figures found from a haphazard collection of tests. It should be contained in the initial hypothesis which guided the selection of tests at the outset. Its verification, however, can be further checked by three supplementary lines of enquiry.

First, we may consider the *internal* evidence, provided by the test-saturations themselves. Secondly, we may ascertain the correlation of each factor with some *external* criterion, i.e., with a set of independent assessments of the abilities assumed—a precaution urged by Burt from the outset. (cf. 1, p. 52). And thirdly, we may endeavour to confirm at first hand the nature of the mental processes assumed by securing *introspections*. In American work, the first type of evidence is apparently deemed sufficient. Quite recently, Dr. Eysenck has urged afresh the importance of checking the results of factorial analyses by an external criterion. In what he calls 'criterion analysis' the factor saturations are correlated with the criterion correlations. But this would seem to be a misunderstanding of Burt's procedure. What should be correlated are not the factor-*saturations* but the factor-*measurements*; [2] formulæ are available for calculating that correlation without explicitly computing the factor-measurements for each individual. In the present research external assessments were obtained for certain abilities; but time did not allow any extensive individual introspections to be obtained from the children tested. The tests were given to several students who were good enough to discuss them with the writer. And the results, with other incidental evidence that need not be detailed here, would seem to warrant our assuming that, on the whole, the tests may reasonably be expected to represent those abilities which, in accordance with the guiding hypothesis, they were intended to represent.

Let us therefore, in the light of the evidence available, briefly examine the nature of the various factors found.

1. *The General or Basic Factor.* With both sets of children, and with each form of analysis employed, a large general factor appears, common to all the thirteen tests. This fully confirms the hypothesis of a dominant cognitive factor of wide generality, entering into non-mathematical subjects as well as mathematical.

How, then, are we to explain the fact that so many psychologists and educationists still maintain that "mathematical proficiency depends on a highly specific ability, and has little or nothing to do with 'general intelligence'"? (i) Earlier writers relied largely on the fact that at times "the mathematical prodigy may be illiterate or even densely stupid." [3] (ii) Those who supported the same

[1] Burt, C., 'Subdivided Factors,' this *Journal*, II, 1949, pp. 41–63.

[2] See, for example, this *Journal*, III, 1950, pp. 124f. and refs.

[3] Mitchell, F. D., 'Mathematical Prodigies,' *Amer. J. Psych.*, XVIII, 1907, 'Summary,' p. 131. Katz says "any ordinarily intelligent child can master the (so-called) 'mathematics' of the school, even to the secondary stage, whereas *higher* mathematics demands special ability" (Katz, D., *Psychologie und Mathematischer Unterricht*). And Tredgold, describing mental defectives committed to institutions, declares that "a few of these persons have an extraordinary capacity for arithmetic and calculation" (*Mental Deficiency*, 1914, p. 307, chapter on *Idiots Savants*).

conclusion by correlational evidence, such as Thorndike and his collaborators, calculated their correlations, not from complete age-groups, but from grades ; this naturally reduces the amount of correlation due to intelligence. (iii) Later investigators admitted correlated factors. Thurstone, for example, in his study of ' primary mental abilities,' can find no " general common factor in the 56 tests analysed," but only a number of specialized group factors. Yet all his group factors prove to be correlated with one another : so that a " general common factor " is really there, though he has declined to extract it. At Burt's suggestion, H. J. Eysenck re-analysed Thurstone's table by the group-factor method. " On applying Burt's very simple procedure," he finds " a general factor accountable for more of the total variance than any other " ; " the disappearance " of this factor, as he says, " is plainly an inevitable result of the method used : no general factor could survive such a procedure." [1]

The first or general factor appears rather heavily weighted by the various mathematical tests. With the summational method, that may be partly due to the large number of mathematical tests ; but the same feature appears in the basic factor of the group factor-matrix. In any case the result is fully in keeping with the common view of teachers, namely, that when the conditions of teaching are approximately the same, then progress in mathematical subjects affords one of the best indications of innate all-round ability.

2. *Special Factors.* (a) *The Mathematical Factor.* With both the boys and the girls the evidence of the bipolar analysis strongly supports the assumption of a broad factor common to the mathematical subjects as contrasted with the non-mathematical.

Three lines of interpretation seem possible. We can attribute such specialized factors to differences either (i) in the experiences or occasions of *learning*, especially in the teaching of the mathematical and non-mathematical subjects at the schools attended by the children tested, or (ii) in the *formal* nature of the mental processes concerned, or (iii) in their *content*.

(i) As regards the first suggestion the following points may be urged : (*a*) School conditions (syllabuses, time-tables, teaching methods, and the like) undoubtedly have more influence on mathematics than on any other school subject ; (*b*) these conditions cannot by themselves altogether account for the factors found : for much the same factors reappear at different schools despite wide divergences in their treatment of the subject ; and finally (*c*) children's introspections show that their own individual ways of learning are far less dependent on the methods of teaching than is commonly supposed.

(ii) Nearly all investigators seem agreed that, if there is a mathematical factor or a number factor, it is not itself a ' formal ' factor (in the psychological sense of the term), i.e., it involves no new type of *process* and represents no specific mode of mental activity found in mathematical work and in no other.[2] Those who maintain the contrary appear to have been misled by the fact that, among the ordinary school subjects, mathematics is the only one that constantly requires the pupil to follow formal proofs. In mathematics, the reasoning, as we have seen, is predominantly of the type known as ' functional deduction ' [3] ; but even this is not distinctive of mathematics as such. Formal proof, functional deduction, and even formal symbols are used quite frequently in psychological tests of reasoning (like those adopted for the present enquiry) ; and, though mathematical tests nearly always prove to be highly saturated with the reasoning factor as thus tested, this factor in its turn often shows equally high saturations for verbal and other non-mathematical problems.

(iii) On these grounds, therefore, Burt maintains that the distinguishing characteristic must lie in the subject matter of the mathematical tests. This seems implied by the other factors that usually emerge, and with which the mathematical factor is contrasted. The difference between it and the verbal and spatial factors is essentially a difference in the mental content, not in the mental operations or form of activity required. What, then, is the specific nature of this content ? " To the psychologist it scarcely seems sufficient for factorists to declare that the content of the number factor is ' numbers ' and that of the verbal factor is ' words.' A ' number ' is not an elementary or simple mental unit any more than a ' word ' is. Words certainly are related to the distinctive human power of speech ; but we cannot suppose that biological evolution has equipped us with a further power for using and understanding numbers." Now, as we have seen (p. 142), the distinctive peculiarity of numbers (as of

[1] ' Primary Mental Abilities,' *Brit. J. Educ. Psychol.*, IX, 1939, p. 272.

[2] This seems confirmed by the work of Jenkins and Jayasuriya, who both failed to find any factor corresponding to the formal functions entering into mathematical problems (such as educing relations of ' correspondence ' or ' order '), and noted that differences in the ' concepts ' and the ' material ' concerned appeared to be far more influential than differences in the mental operations (cf. Jayasuriya, J. E., ' Factor Analysis of Tests of Functional Thinking ' : unpublished Thesis. Univ. Lond. Lib., 1949).

[3] The phrase is taken by Burt from Johnson, W. E., *Logic.*, II, 1922, pp. 123f. In any case, so Burt contends, ' functional deduction ' is not a distinctive *psychological* operation ; it consists merely in applying ordinary types of relational procedures to functional variables. However, the point perhaps deserves fuller investigation.

certain other concepts used in mathematics) consists in the fact that they are *formal variables* ; [1] and this must constitute a psychological as well as a logical distinction. A mathematical ' expression ' is an abstract scheme : it is not the ' expression ' of anything concrete. It is a grin, without a cat. And arguing about schemes, formulæ, or expressions that are entirely schematic, is altogether different from arguing about verbal expressions or about concrete objects : (cf. 16, pp. 436f.). What makes mathematical thinking different from other modes of thought, therefore, is a difference, not so much in the relations to be analysed out or re-applied, but rather in the *Gestaltqualitäten* that have to be apprehended and compared. But, although from a logical standpoint these *Gestaltqualitäten* (like their constituent relations) are ' formal ' characteristics, we must not make the mistake of supposing that they are ' formal ' processes from a psychological standpoint ; they imply special mental contents, not special mental processes.

(b) *The Subdivision of the Mathematical Factor.* As is obvious from the table printed above, the results of factorizing the data here obtained strongly confirm the view that " the factor underlying arithmetical or mathematical tasks is highly complex," and that " with older children of secondary school level there is a broad distinction between an ' arithmetical ' or computational factor, depending mainly on the mechanical use of memorized numerical *associations*, and a ' mathematical factor,' involving the ability to perceive, attend to, and combine numerical *relations* " (20, pp. 186–7). Moreover, the rotated factors suggest that proficiency in geometry depends on the spatial factor far more than arithmetic and algebra. As a result arithmetic and algebra seem to be far more closely related to each other than either is with geometry. Indeed, the correlation table strongly supports the view that a good performance in elementary problem arithmetic may itself be used as an indicator of capacity to make good progress in algebra.[2]

Probably the actual composition of what is broadly termed mathematical ability differs with different individuals. Some light is thrown on this by the sex differences. As will be seen from the table, there are several minor but suggestive differences between the saturations for the girls and the boys respectively. With the girls, for example, the test of mechanical arithmetic is classed, not with the other three mathematical subjects, but with what may be called the number tests ; on the other hand, memory for nonsense syllables and the letter-sequence test are classed with the verbal tests. With the boys the corresponding groups seem to depend more on memory and on reasoning respectively. The negative saturations are also instructive. It would be a mistake to regard them as the artificial results of an incomplete rotation. Even with a graphical procedure it is impossible to suppress them. They suggest that, when other influences have been eliminated, a preference for working with numerical or literal symbols is definitely inimical to working with spatial or geometrical concepts, and vice versa—a difference perhaps attributable in part to a difference in mental imagery. Further, there appears also to be a further antithesis or interference between a preference for routine work and a preference for (or excellence at) problems or riders. The absence of Burt's broader ' intellectual factor ' (covering *both* verbal *and* arithmetical work—Vernon's *v. : ed.*) seems due to the selection of grammar school children only for purposes of this research.

It may be added that his contention that the so-called ' number factor ' is not a simple ' unitary ' or ' primary ' factor but is highly complex has already been endorsed by several recent writers on more general grounds. Hamley, for example, unhesitatingly adopts the notion of ' sub-divided factors ' : " mathematical ability," he says, " is probably a composite of general intelligence, visual and other imagery, ability to perceive and retain certain configurations, etc. : each of these could be further subdivided into simpler factors or components " [3] Jenkins evidently takes a similar

[1] Burt observes that, " since formal variables may be represented by literal symbols or by spatial co-ordinates, just as well as by numerical figures, the characteristic content is not limited to number. But to apprehend that content very much more than forming associations or discerning relations is required " (12). In a more recent paper (cited above) Coombs comes to the conclusion that tests highly saturated with a number factor " really measure some more fundamental process than mere number manipulations *per se* " ; but he appears to hold that this factor simply " represents the rapidity with which well-established associations may be recalled and manipulated " (the associations presumably being associations between symbols and their meaning).

[2] Although both arithmetic and geometry have attracted the attention of factorists, little attention has been paid to algebra : (see Blackwell's list of researches quoted above). The conclusion in the text is borne out by the practical experience of Ruch and Segel : " special aptitude tests are even more highly predictive of success in algebra than general intelligence tests : and arithmetic tests predict as well as, if not better than, the special aptitude tests " (*Encyclopædia of Vocational Guidance*, p. 730).

[3] *The Testing of Intelligence*, 1935, p. 28. Cf. also *id.*, ' Functional and Relational Thinking in Mathematics,' *9th Yearbook, Nat. Counc. Teach. Math.*, 1934.

view, since he describes the mathematical factor as consisting of " a hierarchy of abilities." [1] But there is unquestionably an urgent need for further research, preferably by a combination of the case-study and the statistical methods, to determine the detailed nature of the elementary abilities or sub-factors that enter into mathematical work.

B. The Influence of Character Qualities. Psychologists whose interest lies chiefly in the factor analysis of cognitive abilities are frequently charged with neglecting the influence of motivation. Burt, however, early found that arithmetical and mathematical ability was markedly influenced by emotional and moral traits.[2] For purposes of the present enquiry the traits that seem most relevant are (i) emotional stability as contrasted with instability, (ii) integrated character as contrasted with dissociated or introverted personality, (iii) industry as contrasted with laziness or lack of perseverance. So far as possible, assessments were obtained independently for every pupil from two teachers. The reliability coefficients (as is usual for such ratings) were decidedly poor, namely, stability, 0·58 ; integration, 0·45 ; industry, 0·62. The assessments showed fairly high inter-correlations (0·37 to 0·59) : this suggests that there is either a marked ' halo ' effect or a marked general factor underlying all three.

The correlations with ability tests and school subjects were positive, but never high. The largest was that between industry and the marks for the mathematical subjects (0·45). With the ability tests the correlations were much lower (0·15 to 0·34). Mechanical computation and memory for numbers showed moderately large correlations with emotional stability (0·33), and the tests of mathematical knowledge with industry (0·34) : (for further details see 24). Thus, arithmetic and mathematical performances are undoubtedly influenced by emotional and moral characteristics, but this influence appears small as compared with that of cognitive abilities.

C. The Influence of Environment. Mathematical ability is a subject that is constantly cited by the non-mathematical as an obvious instance of an innate talent or gift. Mathematicians, on the other hand, commonly assure us that " it is only another kind of language ; so that, if children can learn English, Latin, or French, they can learn equally well to answer questions about number or solve propositions by algebra, provided they have an equally strong motive." And much the same view seems implied by the Ministry of Education's regulations.[3] Most psychologists apparently hold that both innate and environmental influence contribute ; but, when we enquire what is the relative importance of the two, the most diverse opinions are encountered.

The view that mathematical ability is innate rather than acquired has usually been advanced to explain cases of *high* mathematical ability (e.g., that of the so-called ' prodigy ') and to support the division of many public schools into classical and mathematical ' sides.' The teacher in the ordinary school more frequently complains of *deficient* capacity for mathematical subjects ; and most of them firmly believe that such incapacities are innate. In either case, if progress in mathematical subjects is really affected by inborn aptitude, then it would plainly be desirable to assess that aptitude in advance, before subjecting a pupil to a curriculum in which he may be foredoomed to fail.

Unlike the faculty for language, the ' faculty for number ' does not appear among the lists of the older faculty psychologists. The earliest arguments for its addition seem to be those of Gall, who supported his proposal by first-hand studies of Colburn and other cases of exceptional calculating

[1] Jenkins, J. W., ' An Analysis of Factors entering into Mathematical Tests, based on the Logical Principles of Mathematics ' (Thesis, Univ. Lond. Lib., 1939) : reference may also be made to the theses of Wheeler, Bennett, Blackwell, Cooke, Dasgupta, and others.

It is interesting to observe that the investigation, published by Mr. Sutherland since my thesis was completed, reveals exactly the same factors as my own. He obtains (Table III) (i) a general, (ii) a number, (iii) a space, (iv) a verbal, and finally (v) a reasoning factor which, like the general factor, has positive saturations for all the tests except mechanical arithmetic. Since the experiment was carried out with 9-year-old boys who had " passed the fifth year in a primary school," it includes no tests for the higher branches of school mathematics ; but nearly all the arithmetical abilities that were tested include at least four factors, and thus appear highly complex.

[2] Cf. ' The Unstable Child,' *Child Study*, X, 1917, pp. 66f., *The Backward Child*, 1937, pp. 545f., and the study of children backward in arithmetic, carried out by Burt with E. Wheeler and other research students (see above, pp. 142-3).

[3] Cf. Hogben, L., *Mathematics for the Million*, p. 1 ; Board of Education, *Handbook of Suggestions*, p. 37 (" Every normal child must acquire the power of speaking, reading, and writing his own language, and some knowledge of arithmetic ").

ability in persons who were otherwise dull.[1] Carpenter supplements such instances by evidence from teachers in London schools. He concludes in favour of " a congenital aptitude for recognizing the relations of numbers " ; [2] and his authority seems to have influenced educational writers like Sully and Lloyd Morgan. On the Continent the chief supporter of this view was Möbius, who published a detailed monograph devoted to the problem.[3]

The earliest statistical study is that of Galton in *Hereditary Genius :* as regards ' mathematical genius ' the data he obtained seemed to him negative or at least inconclusive,[4] though his followers have reached more positive conclusions. The first attempt to apply a rigorous correlational technique to such problems was that of J. M. Cattell and Wissler.[5] In their study of the ' elements of general and specific ability,' Wissler reports (among about 200 students) an average correlation between mathematics and other subjects amounting to 0·54, almost identical with the average of each subject with the rest (0·55). His inference is that, though there may be some evidence for general ability, there is little or none for specific abilities : " whatever it is that makes for correlation in general, class standing seems to hold for *all* the courses." On the whole, therefore, the negative results of Cattell and Wissler appeared to support Galton's view. Burt, however, in his early investigations for the London County Council, relying on clinical evidence quite as much as on correlational, concluded that achievements in mathematical subjects could be limited or favoured by specific capacity as well as by general intelligence, and by innate endowment as well as by environmental influence. Nevertheless, he too stresses the influence of out-of-school interests and occupations, as well as that of formal teaching in the classroom : (1, 12, 13).

The associationist school, naturally enough, rejected all doctrines involving specific faculties. The clearest statement is that of Bain [6] ; and his views seemed at first confirmed by the correlational work of both Thorndike and Spearman. In his first article (*loc. cit.*, p. 275), Spearman gives correlations between school subjects, and reaches the same conclusion as Wissler : all correlations can be explained by ' general ability ' ; there is no evidence for anything like a mathematical faculty or factor. In later writings,[7] as we have seen, he

[1] Gall, F. J., *Organologie*, 1822, Chap. XVIII, pp. 84f. (tr. *The Functions of the Brain*, 1835). The lists of faculties published by Reid and Dugald Stewart, from whom Gall largely took his own list, contain no mention of a mathematical faculty.

[2] Carpenter, W. B., *Principles of Mental Physiology* (1888, an expansion of his earlier ' Outline of Psychology,' 1855, written while professor at University College), p. 231. Among modern medical writers, Dr. Bernard Hollander, quoting with approval S. E. Henschen's paper in *Brain* (1926), concludes that ' a special centre in the frontal lobes ' exists for ' calculating ability,' lying near to those for language and music, but separate from them (*Brain, Mind, and the External Signs of Intelligence*, Allen & Unwin, 1931, p. 212). In a paper entitled ' Congenital Arithmetic Disability and Acalculia ' (*Brit. J. Med. Psychol.*, XVI, 1937, pp. 18f.), E. Guttmann has summarized the alleged evidence for innate disability in arithmetical calculation, resulting from structural or functional disturbances of the brain, and infers that " there can be little doubt that focal brain lesions may produce ' acalculia.' " However, most psychologists, and most competent neurologists, would, it may be imagined, consider the arguments as inconclusive as those for the corresponding condition of ' congenital alexia.' No doubt certain brain lesions may produce difficulties in calculation : some, for instance, are reported to have been followed by a specific loss of memory for the addition and multiplication tables ; but this of itself is no proof of an innate or specific centre for ' arithmetic.'

[3] Möbius, P. J., *Uber die Anlage zur Mathematik*, Leipzig, 1900. Möbius is the writer whom Spearman has chiefly in mind in his attack on the existence of a special mathematical ability. He describes Möbius's view as that " usually accepted " at the time he wrote : *Amer. J. Psychol.*, XV, 1904, p. 276.

[4] Galton notes that, except for the Bernouillis, the names of the most famous mathematicians are absent from his appendix. Nor does he find the evidence he expected in studying the families of the senior wranglers at Cambridge. But he did not, it would seem, intend this lack of evidence to be taken as definite disproof (see p. 198 *ad fin.* : see also *English Men of Science : Their Nature and Nurture*, 1874). In *Inquiries into Human Faculty* (1883), he refers to the ' feats in mental arithmetic of the Bidders,' and adds : " the faculty is hereditary." Bramwell therefore seems to overstate the matter when he writes that " Galton concluded that ' great mathematicians transmit neither general ability nor specialized ability.' " Bramwell himself quotes a number of more recent pedigrees contradicting this conclusion (*Eugen. Rev.*, XXXIX, 1948, p. 152).

[5] ' Mental and Physical Tests,' *Psych. Mon. Sup.*, III, 1901, p. 36. Pearson's early correlational studies of ' mental inheritance ' did not include mathematical ability as such.

[6] Cf. Bain, A., *Senses and Intellect*, 1864, pp. 585f. He argues that the ordinary laws of association suffice to explain the processes of ' constructive ' as well as of ' reproductive ' association ; and he specifically cites the ' mathematical sciences ' to show that " the mental forces involved are no other than those found elsewhere." " The *products* may be different " ; but that, like other forms of specialization, is due solely to a " fascination for the peculiar class of ideas concerned " (pp. 608f.).

[7] Cf. *The Abilities of Man*, 1926, p. 242.

appears to admit, a little reluctantly, some evidence for such a factor, but attributes it to " past experience rather than native aptitude."

Among the investigations carried out under the inspiration of Thorndike by far the most extensive was that of J. M. Rice, who organized tests of 6,000 children from 18 different schools in 7 different cities in the United States. He compares the results obtained for arithmetic with such conditions as (i) home circumstances, (ii) size of class, (iii) hours devoted to arithmetic in the time-table, (iv) teaching methods, (v) amount of homework, etc. He concludes that home environment has little or nothing to do with progress in arithmetic (" the school drawing from the poorest conditions obtained a better standing than any of the so-called aristocratic schools ") ; and considers that the actual supervision of classroom work by means of regular tests or exercises in arithmetical processes is far more important than other conditions so often cited –personality of teacher, time allotted, and the like.[1] Cobb reports a study dealing specifically with the inheritance of arithmetical abilities (4) ; but it is avowedly preliminary and inconclusive. Thurstone, however, as we have already seen, is strongly of the view that, *prima facie* at least, there is " some evidence for the genetic interpretation of number facility as an inherited trait " (14, pp. 82–3). But on the whole most recent writers seem clearly to favour the dominant importance of environmental influences, particularly the type of teaching received at school, though very few quote first-hand data in support of their views.

A Preliminary Enquiry into Environmental Conditions. Apart from the early enquiry by Rice, hardly any psychologists appear to have attempted a direct investigation on the influence of school teaching or of home background. It was therefore considered desirable to determine how far the variations in the pupils' total performances in the tests described above appeared to be affected by differences in conditions of this kind, particularly as, owing to the recent reorganization of the curricula for grammar schools, teaching methods and teaching efficiency vary appreciably from one school to another, and particularly from boys' departments to girls'. To obtain some indication of home conditions, the pupils were classified into five main groups according to the occupations of their parents.[2] An analysis of variance was then carried out to test the significance of variations in the mean performance of the children when classified by (i) school, (ii) department, and (iii) home conditions : for details I may again refer to the work already cited (24). The following conclusions were drawn :

(i) There is a significant difference between the performances of the children in the highest and lowest occupational classes, but otherwise social differences have a smaller influence than most writers apparently assume. So far as it goes, this confirms the conclusion reached by J. M. Rice.

(ii) There is a marked difference between average mathematical attainments in the girls' and boys' departments, the girls being poorer. Among all the tests used in the main investigation, the largest sex differences were found in the mathematical tests—notably computation, arithmetical problems, and algebra. In general intelligence there was no significant sex difference. Moreover, in the tests of special abilities, and (so far as can be judged from the assessments) in motivation, the differences were too small to account for the differences in attainments. Hence, even if part of the sex-difference in attainments is attributable to innate sex-differences in aptitude, most of it would appear due to difference in teaching efficiency. This inference was confirmed by inspectors who were familiar with the schools at which the pupils had been taught (not necessarily those they were now attending) : as one of them puts it, " women are better in teaching English subjects and men in teaching mathematical, though there are many exceptions to this generalization." [3]

(iii) Significant differences were also found between the several schools, quite apart from the division between the sexes. Contrary to Rice's conclusion, performances at schools in good neighbourhoods were better than at schools in poorer neighbourhoods, although there was little difference in general ability. It appeared more conspicuous in mathematics than in elementary arithmetic, but not so great as is commonly observed in literary or cultural subjects.

Influence of Genetic Conditions. The foregoing differences relate to averages only. In one and the same school and department there are numerous individual cases of children, coming from similar homes and possessing equal degrees of general intelligence and (so far

[1] ' Causes of Success and Failure in Arithmetic,' *Forum*, XXXIV, 1902, pp. 437–52.
[2] The classification adopted was a simplified version of that employed by Burt in similar enquiries : he has shown that the principle ' all good features tend to be correlated ' " holds true of environment as well as personality " (' The Assessment of Personality,' *Egypt. J. Psychol.*, II, 1946, p. 6), and that occupation is the characteristic having the highest saturation with the general factor underlying environmental assessments : (see *The Backward Child*, pp. 119f. and refs.).
[3] On this see also Burt, *The Backward Child*, p. 114.

as can be judged) of motivation, who show wide differences in mathematical ability. Accordingly it would seem impossible to explain *all* the differences observed as a result of environmental conditions only. No doubt, had it been possible to secure intensive case-histories of the more conspicuous instances of high or low ability, a complicated interplay of numerous causal influences might have been revealed. But that was impossible. Nevertheless, it may be helpful to conclude by enquiring what general lines of evidence are available to show whether, in part at least, the so-called mathematical factor may not have at least some genetic basis.

" The issue " (as Burt has pointed out) " has almost always been far too simply conceived." Those who favour heredity seem to have in mind some elementary or unitary factor for number-work as such ; and their critics have little difficulty in discrediting this rather naïve notion. " An ' ability ' for ' mathematics,' like an ' ability for music,' is " (so he contends) " hardly the kind of trait we should expect to be produced by biological evolution. And the genetic constitution of the individual can operate only as one underlying condition that favours or that hinders the gradual development and acquisition of mathematical skill or knowledge. Moreover, this underlying condition must itself be a resultant of numerous and varying influences, including temperamental tendencies as well as cognitive."

In support of a genetic basis he enumerates the following points (9, 12, 20) : (i) There are now available numerous pedigrees, indicating that *high* mathematical ability often runs in families. Those of the Bernouillis and the Bidders are the best known ; but those of Cambridge wranglers are still more striking.[1] By itself such evidence is scarcely convincing, since encouragement, tradition, emulation, and opportunity may have played a part. Burt, however, quotes, from his own school cases, several " infant prodigies of a less remarkable type," where further enquiry revealed instances of exceptional computational ability among immediate and remoter relatives. (ii) The pedigrees for children of *low* mathematical ability are even less easy to account for in terms of environmental influence. The case-histories of pupils specifically backward in arithmetic not infrequently include several instances of the same type in the same family—the number being nearly three times as large as that found in the histories of ordinary pupils. (iii) Several writers report positive *correlations* between the results of arithmetic tests for members of the same family (usually 0·3 to 0·5). However, these writers seldom allow for the influence of general intelligence. Using an achievement ratio (arithmetic age ÷ intelligence) Burt and Hewlett report correlations amounting to 0·37 between siblings and 0·28 between parents and children. (iv) In a study of London *twins* the following figures were obtained for achievement ratios in arithmetic : (*a*) unrelated pairs, each pair coming from the same school and from similar homes, 0·18 ; (*b*) siblings 0·37 ; (*c*) non-identical twins 0·38 ; (*d*) identical twins, aged 7 to 9 inclusive, 0·76 ; (*e*) identical twins, aged 11 to 13 inclusive, 0·61. It will be noted that with the twins the resemblance diminished with age ; and this, *prima facie*, seems inconsistent with the view that it was produced by the cumulative effects of post-natal conditions. Before intelligence had been eliminated by partial correlation, the resemblance was found to be greatest in written problems ; after its elimination, the resemblance appeared larger in work of a more mechanical kind. In several cases it seemed related to resemblances in elementary abilities (particularly mechanical memory) and in temperamental characteristics. (v) Observers in many different countries have reported *sex-differences* in mathematical ability, mainly [2] in favour of the boys. As we have just seen, much of this can be attributed to differences in teaching ; even in mixed schools, where the two sexes have been taught side by side by the same teacher, the boys have other interests or ambitions to aid them in their mathematical work. Nevertheless, when all allowance has been made, definite sex-differences still remain which can hardly be explained away in this fashion. Part of it would seem to be the result of different emotional attitudes towards such subjects, and so indirectly related to temperamental differences. (vi) The remarkable *precocity* [3] of many instances of high mathematical ability strongly suggests the influence of a genetic condition : Gauss tells us he corrected a mistake in his father's arithmetic at the age of 3 ; Ampère showed evidence of high calculating ability at about the same age ; F. D. Mitchell (a psychologist who has made a detailed study of his own case) at the age of 4 ; Colborn and Archbishop Whately at 5 ; Inaudi, Safford, and Bidder at 6. Burt cites numerous examples reported by London teachers

[1] E.g., the Turnbull-Hudson pedigree which shows eight wranglers at Cambridge and one 1st class in mathematics at Oxford in two generations. *Eugen. Rev.*, XXXIX, 1948, p. 152.

[2] Not always in every branch : see, for example, *Mental and Scholastic Tests*, p. 326. In view of the poor arithmetical performances of girls at the examination for L.C.C. junior county scholarships, the Education Officer (Sir Robert Blair) obtained data from mixed schools both in London and in Scotland, where boys and girls were taught side by side. The boys still showed some slight superiority (11).

[3] This is one of the commonest arguments in books on educational psychology : e.g., Carroll, H. A., *Genius in the Making* (McGraw Hill Series in Education) : " the fact that the gift appears at a very early age indicates that it is based on some specialized inherited mechanism " (p. 195).

where the ability appeared in the infant school and persisted throughout school life. Of these the majority were instances of exceptional skill in counting or elementary computation ; on the other hand, efficiency in work involving the application of arithmetical rules, and, still more, algebraic reasoning, appeared to depend chiefly on school teaching (more so, according to Burt, than efficiency in any other subject of the curriculum).

The conclusion finally drawn is that " whereas genetic influences appear to account for at least 75 per cent. of the individual differences in general intelligence, they would seem to account for only about 25 per cent. of the individual differences in the specific factor-measurements for arithmetical ability. . . . It is chiefly in exceptional cases of high mathematical ability and of marked arithmetical disability that this factor has to be borne in mind."

Further evidence is still needed. But one point emerges most plainly from the whole of the foregoing discussions. The investigations carried out in the past have been undertaken on too rough and ready a basis. Whether the issue to be studied is the relative influence of genetic factors, or whether it is concerned with the more general problem of analysing mathematical ability into its chief components, the precise hypothesis to be tested needs to be clearly formulated before the experimental work is planned. The ideal method, it would seem, should be based on Burt's method of *combining case-studies with statistical analysis :* a mere statistical analysis is difficult to interpret psychologically, and case-studies by themselves are inconclusive. Moreover, more extensive investigations are required by means of larger batteries, which should include tests, not only of mathematical processes themselves, but also of the numerous simpler abilities related to, or apparently underlying, mathematical processes ; and the mathematical tests themselves should not be merely borrowed from the exercises in current school text-books, but based on modern views of mathematical procedure.

If meanwhile we provisionally accept the view that a 'mathematical factor,' however complex, in some sense or other exists, and that individual differences in that factor are in some appreciable degree innate, then it may, I think, fairly be claimed that tests of the type here found to yield high saturations will prove to be of real prognostic use, and may enable the teacher to assess its amount even before the pupil has reached the secondary school stage. The results therefore have a practical as well as a theoretical value.

IV. SUMMARY AND CONCLUSIONS

1. The conclusions reached by previous writers regarding ability in mathematics are in sharp disagreement. This seems largely due to the fact that each writer has trusted *either* to psychological analysis *or* to statistical analysis, without ever combining the two modes of approach. Accordingly, in addition to presenting tables of factor saturations, the foregoing article attempts to summarize all the existing evidence on the subject.

2. In the main statistical research, 320 pupils, aged $13\frac{1}{2}$–$14\frac{1}{2}$, were tested at four different grammar schools. New tests were constructed for attainments in those branches of mathematics that form part of the grammar school curriculum ; tests for the most closely related abilities were also employed, and assessments were procured from teachers for the character qualities most likely to be relevant. Both in selecting test-problems and in interpreting the results, a special endeavour was made to bear in mind modern views on the nature of mathematics as a branch of formal logic, which have been ignored by most previous writers.

3. The data obtained were analysed (*a*) by the bipolar methods of simple summation and subdivided factors, and (*b*) by the method of group factors, both non-overlapping and rotated. Burt's method of simple summation was found superior to Thurstone's centroid method, and his arithmetical method of rotation to the graphic. It is emphasized, however, that the psychological interpretation of the factors found cannot be satisfactorily derived from an inspection of the factor saturations alone, but must also take into account evidence available from all sources, whether statistical or non-statistical.

4. The bipolar analysis suggested four dominant factors : (i) a general factor, roughly identifiable with general intelligence, and common to all the tests, (ii) a group factor for mathematical ability, common only to the tests involving numerical or mathematical processes, (iii) a verbal, and (iv) a visuo-spatial factor. There was strong evidence that the mathematical factor may itself be subdivided into narrower factors of a simpler type, notably a sub-factor for mechanical arithmetic (closely related to memory) and another for mathematical work (closely related to the manipulation of schemes and relations) ; and there seemed also to be a cross division distinguishing geometry (which appeared to possess a high saturation with the spatial factor, probably in the form of visual imagery) from problem arithmetic and algebra (both of which seemed to depend on facility for dealing with formal variables).

5. The so-called ' number factor ' is thus by no means unitary. Neither is it by any means wholly innate. Ability and disability in mathematical subjects appear to be influenced, perhaps even more than other subjects of the curriculum, by school teaching, and (in a much smaller degree) by other environmental conditions. Nevertheless, they seem also to be partly dependent on genetic tendencies, temperamental as well as cognitive.[1]

REFERENCES

1. Burt, C. (1917). *Distribution and Relations of Educational Abilities.* London : P. S. King.
2. Russell, B. (1919). *An Introduction to Mathematical Philosophy.* London : Allen & Unwin.
3. Whitehead, A. N. (1919). *An Introduction to Mathematics.* London : Williams and Norgate.
4. Cobb, Margaret V. (1919). ' A preliminary study of the inheritance of arithmetical abilities.' *J. Educ. Psychol.*, X, pp. 44–8.
5. Thorndike, E. L. (1922). *The Psychology of Arithmetic.* New York : Macmillan.
6. Buswell, G. T., and Judd, C. H. (1925). *Summary of Investigations Relating to Arithmetic.* Chicago : University Press.
7. Burt, C. (1926). *The Measurement of Mental Capacities.* Edinburgh : Oliver & Boyd.
8. Spearman, C. (1927). *The Abilities of Man.* London : Macmillan.
9. Wheeler, Elizabeth (1927). ' Backwardness in Arithmetic Among London School Children ' (unpublished thesis : University of London Library).
10. Young, J. W. (1927). *Fundamental Concepts of Algebra and Geometry.* New York : Macmillan.
11. Burt, C. (1928). *L.C.C. Reports on Backwardness in Special Subjects* (*II. Arithmetic*) (unpublished).
12. Burt, C. (1929). *A Psychological Analysis of Mathematical Ability with Special Reference to Selection for Secondary Schools* (unpublished).
13. Burt, C. (1937). *The Backward Child.* London : University of London Press.
14. Thurstone, L. L. (1938). *Primary Mental Abilities.* Chicago : University of Chicago Press.
15. Board of Education (1938). *Report of the Consultative Committee on Secondary Education.* H.M. Stationery Office.
16. Burt, C. (1939). ' The relations of educational abilities.' *Brit. J. Educ. Psychol.*, IX, pp. 45–71.
17. Burt, C. (1940). *The Factors of the Mind.* London : University of London Press.
18. Lindquist, E. F. (1940). *Statistical Analysis in Educational Research.* Cambridge, Mass. : Riverside Press.
19. Thurstone, L. L. (1947). *Multiple Factor Analysis.* Chicago : University Press.
20. Burt, C. (1949). ' The structure of the mind : A Review of the results of factor analysis.' *Brit. J. Educ. Psychol.*, XIX, pp. 100–11, 176–99.
21. Reyburn, H. A., and Raath, M. J. (1949). ' Simple structure : a critical examination.' *Brit. J. Psych., Stat. Sect.*, II, pp. 125–33.
22. Burt, C. (1950). ' Group factor analysis.' *Brit. J. Psych., Stat. Sect.*, III, pp. 40–75.
23. Vernon, P. E. (1950). *The Structure of Human Abilities.* London : Methuen.
24. Barakat, M. K. (1950). *Factors Underlying Mathematical Abilities* (unpublished thesis : University of London Library).
25. Sutherland, J. D. (1951). ' Factor rotation by the method of extended vectors.' *Brit. J. Psych., Stat. Sect.*, IV, pp. 21–30.

[1] I am much indebted to Dr. Charlotte Banks for help both in the statistical analysis and in preparing my manuscript for publication. My thanks are also due to the Headmasters who allowed me to work in their schools and to Professor Burt for permitting me to quote so freely from his reports and memoranda.

Vol. IV
Part III

The British Journal of Psychology,
Statistical Section

November,
1951

INFORMATION THEORY AND INTELLIGENCE TESTS

By W. E. HICK

The Medical Research Council Applied Psychology Unit, Cambridge

I. *The Measurement of Information.* II. *Types of Question.* III. *Information in a Question.* IV. *Information in an Answer.* V. *Multiple-Choice Questions.* VI. *Equalizing Information over Abilities.* VII. *Summary.*

I. THE MEASUREMENT OF INFORMATION

The purpose of this paper is to draw attention to some respects in which the theory of information appears relevant both to the notion of intelligence and to the design of tests. Although the whole process of intelligence-testing has been subjected to such intensive mathematical analysis for so many years that the remarks put forward here are unlikely to embody anything fundamentally new, nevertheless the language and methods of information theory enable certain points to be exhibited in a simple and striking form.

There seem to be at least three varieties of ' information theory ' in existence. These have been called respectively the statistical, the selective, and the structural-metrical (5). Since they are adapted to different uses, there is no question of rivalry between them ; and in many cases they coincide. Fisher's theory of statistical information is well known (1, 4). The theory of structural and metrical information developed by Gabor and MacKay will not be introduced here, although it may well be found to have applications in psychology (5, 6). The so-called selective information, particularly associated with the names of Wiener (2) and Shannon (3), is the most immediately illuminating, and provides the approach followed in this paper.

In this system, the information contributed by an event of probability p is defined as $-\log p$. This measures only the *amount* of information, and says nothing as to its nature, meaning, or relevance to any proposition. These attributes must be established, if necessary, on collateral grounds.

Since $p \leqslant 1$, information is never negative, and is a monotonic decreasing function of p. Thus an almost impossible event is almost infinitely informative, if it occurs ; and an event which is certain to happen conveys no information when it does happen. Further, the contributions of information from independent events can be added to give the total quantity. These are reasonable requirements of a measure of information ; and the logarithmic function is the only one which satisfies them.

If base 2 logarithms are used, the unit is the ' bit,' corresponding to $p = 0{\cdot}5$. However, the choice of base is of no importance in principle ; and, when it happens to be more convenient to work in Napierian or in common logarithms, the writer confesses to a private convention in which ' naps ' and ' decs ' are the respective units.

The prior expectation of information is the average, namely,

$$-p \log p - q \log q,$$

where $q = 1-p$. Generally, if there are a number of alternatives having the probabilities $p_1, p_2, \ldots p_n$, the prior expectation is

$$H = -\sum_1^n p_i \log p_i .$$

Both Shannon and Wiener (*loc. cit.*) have pointed out that this is a form of the entropy expression well known in statistical mechanics ; and we shall follow them in referring

to it by that name. It is a measure of our uncertainty as to what will happen, whereas $-\log p$ measures, in a comparable form, our doubt as to whether the particular event will occur.

In the case of a continuous distribution, the exactly equivalent formula would be

$$H = -\int p(x) \log (p(x)dx) \,.\, dx, \qquad (1)$$

where $p(x)$ is a distribution function and the integration is over the range including all non-zero values of $p(x)$. If dx is regarded as constant over the range, this reduces to

$$H = -\int p(x) \log p(x) \,.\, dx - \log dx. \qquad (2)$$

The second term tends to infinity, on account of the unlimited number of possible values of x. But this infinite term disappears in most practical applications, and therefore the integral alone is taken to represent the entropy (or ' relative entropy ').

But suppose we wish to make the probability of occupance of dx constant over the range of x. We write $p(x)dx = \delta k$, and substituting in (1) we find $H = -\log \delta k$, which is simply the information due to the occurrence of one out of $1/\delta k$ equiprobable events. But we are not specifying x with the same precision over the whole range, because the zone of uncertainty, dx, varies inversely as $p(x)$.

Wiener (*loc. cit.*) applies this to the case of a rectangular distribution, as a simple example of the relation between information and precision. Here $p(x) = p$ (constant) in the interval (a, b) and is zero elsewhere. The area under the rectangle must be unity ; therefore $p = 1/(b - a)$. Substitution in (2) shows that $H = \log (b - a) - \log dx$. Now suppose that further knowledge enables us to say that x lies somewhere—we have no idea where—in the smaller interval (c, d). The entropy is now $\log (d - c) - \log dx$. How much information have we gained from this further knowledge or more exact specification? If we subtract the former entropy from the latter, we obtain

$$\log \frac{(d - c)}{(b - a)}.$$

Note that this is negative, because our currency is ' uncertainty ' or ' expected information,' and our loss in this respect is equal to our gain in actual information.

In discussing communication problems, Shannon employs the following useful scheme. The initial events are *messages* (x) generated by a *source*, which latter can therefore be assigned an entropy $H(x)$. The messages are encoded by a transmitter and despatched along a *channel*. In their course, they may be interfered with by more or less random disturbances—the ' noise ' of the communication engineer. The received signal y belongs to an ensemble whose entropy, $H(y)$, may be greater or less than $H(x)$.

II. TYPES OF QUESTION

Two basic types of question appear to be used in intelligence tests, namely, (*a*) the pure deductive and (*b*) the deductive-inductive. Unless one admits a pure inductive type, there is no other possibility. It will be appreciated that this classification has nothing necessarily to do with the psychological phenomena associated with the act of solution ; the answer may seem to come by a flash of insight or by hard thinking, or by any other way—perhaps just by the subject being the same sort of organism as the psychologist who devised the question. But logically speaking, a correct answer that was not merely a lucky guess must have been arrived at by deduction or induction or both.

The deductive question, if understood, can, in principle, be solved by what we call pure reason, because all the necessary information is given. Since, however, such questions do not correspond to the problems most commonly met with in real life, it is usual to include questions of an inductive or mixed type, from which more or less of the essential information is missing. A question of this kind has no unique solution, in the logical sense. It is generally accepted nowadays that there is, and can be, no *logical* justification whatever for inductive reasoning. It is simply an act of faith. Even the argument that it works in practice is inductive, and is therefore a *petitio principii*. Nevertheless, it happens from time to time that a psychologist imagines that he has discovered a question which is not purely deductive, but which has only one correct answer—i.e., one demonstrably correct answer, as distinct from a solution arbitrarily chosen by the psychologist himself.

The confusion is often excusable. All systems of classification, being by nature artificial, have their Achilles' heel—the borderline case. In reality, no question is either wholly deductive or wholly inductive. The onus is always on the subject to guess what the question means, and thence to infer what the most probably correct answer is. It may be that the emphasis is on reducing the question to what is, for the subject, the canonical form—i.e., disentangling its meaning—or on the logical (deductive) procedure of solution, or on the divination of the answer from a judicious combination of general knowledge and the special information given in the question. Only the last is what we should normally call an inductive question.

In the language of information theory, the subject receives a signal, the question, and from that he is required to reconstruct the original message, the question with its answer. The process which has obliterated the answer, although its agent is the psychologist, is, from the subject's point of view, statistical or probabilistic in nature. It is a form of ' noise '—i.e., irrelevant disturbance—and the correct reconstruction of the message can only be made, except by chance, from a knowledge of the statistical structure of this ' noise.' On the other hand, where there is no true loss of information, the problem, if there *is* a problem, becomes merely one of deciphering according to a known key, that is to say, one of reasoning from the general to the particular ; and that is what I mean by deduction.

In other words, an intelligence test, in the ordinary sense of the term, is a test of knowledge or skill. Clearly such things as ' ability to perceive relations ' come under the heading of knowledge or skill, whether they are acquired by learning or by maturation. Needless to say, ' knowing how ' and ' knowing that ' are here classed together as ' knowing.' The distinction is of practical importance ; but the essential unity of all forms of ' knowing,' under whatever names they masquerade, must be grasped if full use is to be made of information theory. The test subject is, in fact, what is called a predictive filter, designed to compensate for certain deficiencies in the input, and we want to find out how good a filter he is for certain types of input. The word ' predictive ' refers, of course, only to that future which belongs to the subject's frame of reference ; it means simply that the time when he will see the psychologist's version of the answer has not yet arrived ; it may never arrive, but the task is to predict what the answer would be if that time ever did arrive. In ' local time ' what is unknown is always in the future.

III. INFORMATION IN A QUESTION

It would appear that all test questions, whether deductive or inductive, can be *formally* reduced to incomplete number sequences, together with full or partial instructions for completion. Thus, a matrix of numbers can be rearranged, without loss of logical content, in a single row. A diagram, or any object, can be described by a matrix of co-ordinates. A process merely requires the addition of the time dimension to the co-ordinate space. Special relations, such as occur in ' 2 is to 4 as 4 is to x,' are instructions which make the question more deductive in type ; this example is, of course, purely deductive, if the relations are understood in the usual sense of proportionality. On the other hand, ' bull is to cow as dog is to . . . ' only gives partial instructions for solution, because the links ' is to . . . as . . . is to . . . ' do not here conventionally imply proportionality. It is a matter for a consensus of opinion whether they here imply *any* unique relation, and, if so, what it is.

Now, if all questions can be formally reduced to the completion of number sequences, it is natural to ask whether it is theoretically possible to state the amount of information given and required. It would be agreed, for instance, that the completion of 2, 4, 6, x is easier than the completion of 2, 4, x, y, and that the difference is due to one more term being given, rather than to one term less being left out. The procedure, conscious or otherwise, being to infer the rule and thence to deduce the

consequences, the former sequence suggests one probable rule, the infinity of other possible rules being individually and collectively improbable. The latter sequence obviously suggests two probable rules, with little to choose between them. The point is that the information given depends on the probabilities, as seen by the subject, of such rules as happen to occur to him. (A rule which does not happen to occur to him has zero probability of correctness, from this aspect.) Information is a function of the probability of the message or symbol ; it is not a property of the message or symbol by itself. Consequently it is impossible to derive any absolute *a priori* measure of the information given in a question. There must be some assumption concerning the subject's notion of the relative probabilities involved.

It may be remarked here that, although in other fields we may choose to attribute to probability a real objective existence, independently of a particular observer, in information theory we must always refer it to a specified person, real or hypothetical.

The conventions involved in the solution of actual number-sequence problems are of passing interest. Let S_n be the complete sequence, of which $n - r$ terms have been omitted. If the terms are imagined to be values of a polynomial at intervals of an argument x, an $(n-1)$th degree polynomial will, of course, fit S_n perfectly. But the subject can only determine an expression of the $(r-1)$th degree ; or, what amounts to the same thing, he can only take differences up to the $(r-1)$th order. However this alone is no basis for extrapolation. The essential thing about what is called ' being given all the information required ' is that, in fact, the subject must be given rather more. As in science, he must use some of his data in forming his hypothesis and the remainder in testing it. To paraphrase the Bellman : " What I tell you two or more times is true." Given a column of constant or uniformly oscillating differences, this requirement is clearly fulfilled ; the confirmatory evidence may be very weak, but since there is nothing to contradict it, it suffices to tip the scales. In using this method, or in testing for a geometrical progression, the subject also assumes that equal weight is to be attached to all the terms ; which is to say that he believes in Bayes' axiom, in cheerful disregard of all the philosophical controversies about it.

IV. INFORMATION IN AN ANSWER

Although, as has been shown, we cannot speak of the information in a question without regard for the person being informed, once we do admit the empirical element, a number of possibilities arise. Only a few simple examples will be touched on here, since it is for those specially concerned with test design to carry the matter further.

First, it is clear that if the answer to a question must be absolutely right or wrong, the most information is given if the chances are equal. This can be seen by differentiating the entropy $-p \log p - q \log q$ with respect to p (or q) and equating to zero. We now come to the delicate question of what this information means—what it is about. It has nothing directly to do with the estimation of the distribution parameter p, because that is a matter for Fisher's ' statistical information,' which gives quite a different result. Moreover, p is assumed to be known ; i.e., our estimate of it is supposed to be unaffected by the particular event from which we derive information.

In communication problems, information is related to the number of symbols required to represent the message. Thus, a message of n symbols, each chosen from an alphabet containing k symbols, is one of the k^n different messages of that length that are possible. Instead of symbols and an alphabet, we can think of a portfolio of statements ; how many statements, each a choice out of k, must we string together in order to classify the subject with a given degree of precision ? To every degree of precision there corresponds an amount of information, and the fewer statements are needed to represent it, the more information, on the average, must each one contain. Any compound statement can be decomposed into the so-called ' atomic propositions ' —statements that something either is so or is not so, without qualification. If a subject's answer must be either right or wrong, it amounts to an atomic proposition. For our purpose, then, selective information is a measure of the number of atomic propositions required of the subject in order to place him in one of a given set of

classes of the population. For a more detailed discussion, Barnard (8) may be consulted.

The most economical way of identifying one out of n equiprobable things is to proceed, if possible, by equal cuts. This is only possible if n is an integral power of 2 ; but in other cases it appears that the maximum ' extravagance ' need not exceed about 0·09 ' bits.' (This small discrepancy is due to the impossibility, in such a case, of making the ' atomic propositions ' equally likely to be true or false—or of making the halves into which the set of objects is divided at each cut equally likely to contain the object sought.) Hence an intelligence test should, in theory, be a ' branch process ' ; i.e., the first question should have a 0·5 chance of being answered by anyone from the general population. If the subject answers it, the next should have a 0·5 chance of being answered by anyone who has been successful with the first ; and so on. Three questions extract $\log_2 8$ units of information, and therefore allocate the subject unfailingly to his correct 12·5 per cent. group—but only with respect to what he has *done ;* his ability to do it again (or something equivalent) depends on the distribution of fluctuations of ability. There are obvious practical objections to administering a test exactly in this way, although, of course, it is the method approximated to in an informal interview.

To illustrate the approach, let right and wrong answers be denoted respectively by 1 and 0. Then the sequence of answers can be represented as a number in the binary scale, the first answer being indicated by the first digit (reading from the left) ; thus, 1010 means right-wrong-right-wrong for a sequence of four answers in a branch process scheme. Bearing in mind the arbitrary but necessary distinction between real ability and the various chance factors which may have altered the result, we can call the observed binary number the ' signal.' The message which it is desired to reconstruct is then the binary number representing the subject's true ability.

If x represents the ensemble of possible messages and y the signals, the information received is (Shannon, *loc. cit.*) :

$$\begin{aligned} R &= H(x) - H_y(x) \\ &= H(y) - H_x(y) \\ &= H(x) + H(y) - H(x, y), \end{aligned}$$

where $H_y(x)$ is the entropy of x when y is known, $H_x(y)$ is that of y when x is known, and $H(x, y)$ is the joint entropy based on the probabilities of all possible pairs (x, y).

$H(x)$ we know, because, by trial on a sufficient sample of the population, we have made it nearly equal to $n \log 2$, for sequences of n questions. But, no matter which of the other entropies we use, we have to estimate the ' within subject ' probabilities ; and this involves either repeated testing with the same questions or with a succession of equivalent sets.

It may be noted that the observed binary number is the result of a non-Markovian process, since if the first digit is wrong, the second digit is likely to be right, in the sense of tending to correct the signal. If the first two digits are wrong, the third is even more likely to tend towards correcting the signal ; and so on. The probability at any stage depends on all the previous stages. At the worst, the subject may begin by putting himself in the wrong half of the population, and no subsequent answers can get him out of it. As one might expect, reliability, economy of effort, and exactness of statement are in competition with one another.

The information transmitted (R) appears to be rather sensitive to the probability of error in a digit. For example, a chance of error of 0·25 (the same for all digits) obliterates over 80 per cent. of the information. Whereas a chance of error of 0·01 destroys only about 8 per cent.

It might be thought that precision and practicality could be achieved by means of multiple-choice questions arranged in groups, each group being self-scored and therefore indicating the next group to be attempted. Thus, suppose the subject must work through a sequence of k groups of j questions each, with t choices per question. The total number of questions available must be

$$j \sum_{r=0}^{k-1} t^{jr}$$

which is an absurdly large number in any practical case.

V. MULTIPLE-CHOICE QUESTIONS

If t alternative answers are given, t should obviously be not very small, to reduce the risk of a correct guess. If the subject has to find the answer for himself, t is

effectively large, though not infinite. The maximum entropy results if all t choices are equally probable, but the maximum useful information (apart from special assumptions about guessing, as exemplified below) is obtained by equalizing the probabilities of the categories actually used in scoring—namely, ' right ' and ' wrong ' in the simplest case. If this is done, the total entropy per question is $\log 2 + \frac{1}{2}\log(t-1)$, of which the second term represents irrelevant information due to the arbitrary subdivision of the class of wrong answers (assumed equiprobable in the above formula). But if a meaningful subdivision, in terms of degrees or kinds of incorrectness, can be found, a larger quantity of useful information can evidently be extracted.

It has been tacitly assumed in this paper that the questions are of equal relevance, or at least that they can be weighted for relevance. In principle, this problem too can be expressed in terms of entropies. For example, suppose the population can be distributed by some external criterion, such as examination results. It will normally be required to predict the category of the criterion from the test score. Although the former is in the future, it can be regarded as a message which gives rise, with various probabilities, to the various test scores, and from these we wish to reconstruct it. In case the word ' reconstruct ' seems inappropriate to something which has not yet happened, we can easily imagine that the future message is the determinate result of the present state of the universe, and that that is what we are really trying to discover. Given the entropies, which are clearly discoverable in principle, we can calculate the average information transmitted.

Shannon's notions of ' channel capacity ' and ideal coding can be introduced here. In the case of the external criterion, for a given test there will be an optimum way of dividing the range of values of the criterion into categories so as to make the fullest use of the capacity of the test to convey information. The problem is to adjust the probabilities associated with the criterion categories so as to maximize the information received (which may be on a per-test or a per-question basis).

By way of illustration, we may revert to the case of guessing in the multiple-choice question. The simplest possible model we can set up involves two states of the subject—one in which he can answer the question with certainty, and another in which he can only guess at random. There is a chance P of getting a subject who is in the former state (A), and a chance Q of getting one who is in state B. With t alternatives given, the chance of the right answer is $P + Q/t$ and the chance of any particular wrong answer is Q/t.

Of the three expressions for the information received that were given above, $R = H(x) + H(y) - H(x, y)$ is as convenient as any. In general, we have a set of input probabilities P_i (P and Q in this example) and a set of transition probabilities p_{ij} : i.e., p_{ij} is the chance, given input i, of receiving signal j. The entropies are

$$H(x) = -\sum_i P_i \log . P_i$$

$$H(y) = -\sum_{ij} P_i p_{ij} \log \sum_i P_i p_{ij}$$

$$H(x, y) = -\sum_{ij} P_i p_{ij} \log . P_i p_{ij} .$$

In this example, $p_{11} = 1$ and $p_{1j} = 0$ for all j. Also $p_{2j} = 1/t$ for all j. The information received is

$$R = \frac{1}{t}\left[Pt . \log . t - Q(t-1)\log . Q - (Pt - P + 1)\log(Pt - P + 1)\right].$$

Solving $dR/dP = 0$ for $t = 2$, for example, we find the optimum value of P to be 0·6. The information received on the average is 0·32 bits, instead of the one bit which is the theoretical maximum for two choices. The chance of getting the right answer is $P + Q/2 = 0{\cdot}8$, giving a prior uncertainty as to what the subject will actually do of 0·72 bits. The information R has to do with discriminating between the states A and B; the uncertainty about the outcome is greater because there are three joint events—two ways of succeeding and one way of failing.

The optimum value of P is

$$P(\text{opt.}) = \frac{t^{t/(t-1)} - 1}{t + t^{t/(t-1)} - 1}$$

which tends, as it should, to $\frac{1}{2}$ as t becomes large. P can be adjusted to the optimum value by varying

the difficulty of the question until the observed relative frequency of success approximates to $P + Q/t$. Thus, for $t = 3$, $R = 0{\cdot}47$ bits if $P + Q/3 = 0{\cdot}72$, and for $t = 4$, $R = 0{\cdot}56$ bits if $P + Q/4 = 0{\cdot}68$.

It is interesting to note the correspondence between this method and the well-known correction for forced-guessing in a threshold experiment by subtracting the number of wrong answers from the number of right answers and taking the result to be the number of genuine perceptions. With two alternatives, we found the optimum proportion of right answers to be 80 per cent. ; subtracting the remaining 20 per cent. we infer that there would be 60 per cent. of ' genuine perceptions,' which agrees with the value of 0·6 found for P. But clearly the theory of information gives us much more than a mere correction for guessing. Unfortunately, although experiment may show that this simple model will not do in some particular case, it will not tell us what to substitute ; at least one transition probability is necessarily concealed.

VI. EQUALIZING INFORMATION OVER ABILITIES

In the case of an intelligence test administered in the ordinary way, the greatest amount of information is obtained about those subjects in the middle range who have about a 50 to 70 per cent. chance, on the average, of answering the questions correctly. With regard to the group, if it exists, which is *a priori* practically certain to answer all the questions correctly, the test cannot discriminate between its members, and therefore, from that point of view, gives no information ; and the same for those who have no chance of answering any of the questions. But it appears that a proper choice of questions would permit the information to be distributed over the range of ability in any way desired. Let the questions be classified with respect to difficulty (i), and the subjects with respect to ability (j). The range of difficulty is assumed to be sufficient to ensure that, for all abilities, there will be at least one question with practically 100 per cent. chance of being solved, and at least one with practically zero probability.

The information per question will be a function of difficulty and ability, say $F(i, j)$. In the simplest case, $F(i, j) = -p_{ij}\log . p_{ij} - q_{ij}\log . q_{ij}$ (where p_{ij} is the probability of solution, given i and j). But special elaborations, such as that of the previous section, could be used. It is unlikely to be worthwhile to carry the process of refinement very far.

The information at the level of difficulty i is to be weighted by W_i ; in other words, W_i is to be the number or proportion of questions of difficulty i in the test. Thus

$$\sum_i W_i F(i, j) = I_j ,$$

where I_j is the information obtained about subjects of ability j. Provided there are as many values of j as there are of i, we have a set of equations determining the W_i. If we wish to make all the I_j equal to some quantity I, the solution takes the comparatively simple form

$$W_i/I = |A|/|F|$$

in which $|F|$ is the determinant of the coefficients of the W_i and $|A|$ is formed from $|F|$ by substituting units in the ith column.

The above treatment—and, indeed, everything in this paper—is merely intended as a suggestion. There are some obvious practical complications, such as interactions due to time limits and practice improvement in the test. However, no great precision can reasonably be aimed at, and it may appear that the complications are not really serious. The procedure is bound to be laborious, since it means trying a large number of questions on a large number of subjects ; but once done, it might remain useful for a long time.

Finally, I would like to express thanks to Sir Cyril Burt and Dr. A. D. Harris for most valuable suggestions and discussion.

VII. SUMMARY

1. After describing the concepts of 'selective information' and 'entropy' as these terms are used in the theory of information, it is suggested that the process of answering test questions may be broadly interpreted as a communication problem.

2. The principles of coding in order to maximize the transmission of information per question or per test are discussed and illustrated by examples. The case of random guessing in a multiple-choice question is worked out to show the use of the different varieties of entropy.

3. Finally, the basis of a method of designing a test so as to distribute the information obtained uniformly or as desired over the range of abilities has been examined; and it is concluded that this is feasible, though laborious.

REFERENCES

1. Fisher, R. A. (1935). *The Design of Experiments.* London : Oliver and Boyd.
2. Wiener, N. (1948). *Cybernetics.* New York : John Wiley and Sons.
3. Shannon, C., and Weaver, W. (1949). *The Mathematical Theory of Communication.* Urbana : University of Illinois Press.
4. Fisher, R. A. (1950). *Contributions to Mathematical Statistics.* New York : John Wiley and Sons.
5. *Proceedings of the Symposium on Information Theory* (1950). London : Ministry of Supply (in the press).
6. MacKay, D. M. (1950). 'Quantal aspects of scientific information.' *Phil. Mag.*, 7, XLI, 289.
7. Burt, C. (1951). 'Test construction and the scaling of items.' This *Journal*, IV, 95–103.
8. Barnard, G. A. (1951). 'The theory of information.' *J. Roy. Stat. Soc.* (in the press).

Vol. IV
Part III

The British Journal of Psychology,
Statistical Section

November,
1951

A FACTORIZATION OF TESTS OF PERSONALITY SOURCE TRAITS

By RAYMOND B. CATTELL
Department of Psychology, University of Illinois

I. *Definition of Methodological Approach.* II. *Relation to Earlier Experiments.* III. *Results of Factorization.* IV. *Summary. Appendix.*

I. DEFINITION OF METHODOLOGICAL APPROACH

A variety of enterprises in individual, clinical, and social psychology await the development of objective tests of personality. By objective tests are meant tests referring to personality source traits, which shall be as clearly defined as primary abilities, and which in scoring will be as free of ' crystal ball gazing ' as intelligence tests.

The report made here is one of several in a unified research plan prosecuted continuously since 1932 (1) (4) (6) (7) (8), and of which the first complete summary of principles and findings was published in book form in 1946 (5). Briefly the general plan has been to obtain outlines of the chief personality factors first in terms of behaviour ratings (4) (6), then in terms of questionnaire responses (8), and so to use the evidence from these two ' make-shift media ' to design variables whereby personality source traits (7) can finally be stabilized in the third medium—objective tests. Throughout this work the source trait is defined as a factor, obtained by the use of factor analysis ; but certain special conditions are presumed to have been observed, notably (i) the distribution of variables throughout the ' personality sphere ' (5) ; (ii) the use of large correlation matrices ; (iii) the use of blind rotation for simple structure ; (iv) the acceptance of oblique factors ; and (v) demonstration of the reappearance of the factors in analyses using *R*–, *P*– (and possibly *Q*–) techniques (9).

In guiding further work the chief emphasis has been placed on the factorization of data from behaviour ratings (*BR*) or from life records (5). This has the virtue of dealing with behaviour of social utility *in situ*. *BR* factorization has yielded 15 source traits, labelled contingently *A* up to and including *O*, in approximate mean order of variance occurring in the three chief studies which have so far appeared (4) (5) (6). Each of these source traits presents us with a fairly definite hypothesis, or one or two alternative hypotheses, upon which to begin designing objective tests of promising validity for the source trait in question. Thus the rating description of factor *C*, Emotional Stability, suggested that the essential character of the factor is dynamic integration centred in self-sentiment, and that a test of consistency and constancy of attitudes would provide a valid measurement. Accordingly, a test of attitude fluctuation (3) has been devised, as well as a test of internal consistency of attitudes. As previous (3) and present results show, this has proved to be loaded with a factor which has every indication of being the same as the rating factor *C*.

Other illustrations of the direction of test design by hypotheses drawn from *BR* factors may be given from Cyclothymia *vs.* Schizothymia (*A* factor), Surgency *vs.* Desurgency (*F* factor), and Paranoid schizothymia (*L* factor). For the last a test of ' threshold of self-reference ' was devised, while for the first a measure of relative intensity of social to impersonal interests was indicated, as well as a measure of the tendency to remember views agreeing with one's own in contrast to views disagreeing with one's own. This last, the consonant-dissonant opinions ratio, has its germ in the particular conception of personal rigidity revealed in the rating factor. Examination of individuals high and low in the *BR* Surgency *vs.* Desurgency factor suggested early (2) that fluency of association would be related to surgency, while later examination of the corresponding questionnaire factor revealed (8) a central characteristic of happy-go-lucky carelessness, which is being investigated by objective tests measuring frequency of careless errors. That errors in relation to speed are effectively loaded in the surgency (or hysteria) factor has meanwhile been independently confirmed by Eysenck (13).

The essential methodology here is to invent objective tests judged to involve particular forms of behaviour indicated by the rating and questionnaire factors. The

particular behaviour is detected (*a*) by clinical observation of individuals very high and very low in the factor, and (*b*) by abstract reasoning about the essence of a factor, based on the empirical construct presented by the collection of factor loadings. The factorization of objective tests then proceeds in a series of successive related factorizations, in which each new battery of objective tests carries over salient variables—' factor markers '—from the preceding one, while adding new tests calculated to yield still higher loadings in the factors so far found. Research thus proceeds by successive approximations, in a cycle of test-hypothesis-test, the period from hypothesis to hypothesis being in our experiments about two and a half years, i.e., the measurement and subsequent calculation for each battery takes about two years.

II. RELATION TO EARLIER EXPERIMENTS

The first attempt to factorize a battery of objective tests designed to be large enough to cover most known *BR* factors was made in 1945. In the present and related articles we shall refer to this as the ' R_1 study,' to distinguish it from the R_2 and later batteries. In the R_1 study, the first of the series of large battery factorizations, it was felt to be essential to make contact with any adequately indicated factors in previous less extensive objective test researches ; for the stream of factorial research must in general carry with it the representatives of any earlier factor discovery, if one is to avoid the confusion and uncertainty which arises from accumulation of non-overlapping factorizations. A survey published (5) about this time indicated about sixteen such factors, including Thurstone's perceptual factors (16), and such factors as disposition rigidity, ideational fluency, honesty-dependability, dynamic momentum, determined will-character, and oscillation. At least two markers for each of these were included in the 1945 study along with so many new tests, designed on a promising theory (constituting the 13 ' older ' tests, and some newer tentative markers discussed below), thus bringing the total battery up to 90 distinct tests.

These tests were administered, mostly in group form, but about one-sixth as individual tests, to 280 20-year-old women students and 130 20-year-old men students, involving about nine hours of testing per subject. Various labour-saving devices have been invented to permit the handling of factorizations on the scale required by the theoretical and methodological standards of such studies as this (9). The most satisfactory, in general, is the simple one of breaking the battery into two overlapping batteries, since the computing labour is roughly equal to the square of the number of variables, and $2\,(n + h)^2$ is generally less than $4n^2$ (where n is half the number in the battery and h is half the number overlapping the two matrices. However, the overlap must be carefully arranged and, where time permits, it is certainly best to finish the factorization of one part first and to carry over into the second matrix only those variables which are known to be good markers for factors found in the first. This was done in the present research.

The first 48 variables, containing all the 13 markers for earlier historical factors, were factorized together and the results were reported in 1948 (7). This we shall call the ' R_1F_1 study,' meaning the first factorization of the first research. (Actually in the whole decade of research it may be called the ' *OT*, R_1F_1 study,' indicating its belonging to the series of objective test factorizations.) All but four of the 16 factors found in earlier researches were duly located in the new battery, though it was found that some factors considered to be distinct in these earlier, isolated, small-battery studies had actually been different representations of one and the same factor. The present article reports the factorization of the remaining 64 variables, including the 22 markers carried over from the first factorization to represent the 11 factors found therein, which also represent the earlier historical factors. This we shall call the R_1F_2 (or *OT*, R_1F_2) study. The present R_1F_2 battery is thus largely composed of newly developed tests—the exceptions being the 22 markers overlapping with the R_1F_1 study reported in 1948. The latter was largely

made on already existing test designs interspersed with a few measures designed on new principles. Thus the 1948 study was a bow to the past, as the present study is a formal introduction to a future of entirely new test designs.

The tests employed are described in the Appendix. It has been necessary to employ the utmost condensation ; and, despite their significance for psychological theory, they are given in outline form only. Fuller particulars of individual tests than can be given here may be had at cost of time for clerical transcription on application to the author. The 22 marker variables (two for each factor, except where overlap brings it to three), carried over for the 11 factors of the previous factorization, are placed at the beginning of the list, as titles only. Placed after each of these is the number it had in that factorization. These particular variables have not been described afresh here.

The first set of variable numbers in the Appendix below (the first column) represents an overall numbering system or ' master index,' set up and retained throughout the two published studies (R_1F_1 (7) and R_1F_2) and two further studies in the press (R_2F_A (11) and R_2F_N (12)). However, because it is also desirable for tabular convenience to have variables consecutively numbered in the matrices of any one factorization we also record in each article, in addition to the stable index system running through all the studies, a temporary numbering system, here in brackets. The latter is necessary for looking up loading patterns. A translation from one to the other is readily made in one direction from the first column of the list.

Immediately below each description of test and scoring is given a brief rationale of the design, generally in terms of the *BR* factor (5) at which it is aimed. Naturally space prevents credit to the many straws of evidence on which the hypotheses are based.

The overall design of experiment does not call for high reliability for each test. At this exploratory stage there must be many and short tests, from which any test having a good loading relative to its reliability can be selected for lengthening. The reliabilities, such as they are, are set out, corrected by the Spearman-Brown formula to full length, opposite the matrix index number, though it has not been feasible to obtain them for tests of non-divisible type.

TABLE I. UNROTATED VECTORS

Trait	Factor 1	2	3	4	5	6	7	8	9	10	11
1	—·25	·06	·04	·30	·09	·30	·42	·09	·17	·13	·13
2	·18	·00	·09	·02	·19	—·24	—·20	·01	—·15	—·25	—·08
3	·14	·09	·23	·05	·21	·14	·44	—·06	—·06	·10	—·09
4	—·15	—·53	—·13	·11	—·07	·26	—·04	·07	—·02	—·04	·53
5	—·32	·41	—·16	—·45	·05	—·10	·11	·03	—·11	—·03	·16
6	—·17	—·09	—·39	—·08	·03	·08	·00	·09	·30	—·04	—·14
7	·05	—·14	·34	·18	·06	—·23	—·34	—·21	—·24	·32	—·13
8	—·08	—·07	·30	·02	—·15	·31	·28	·06	·57	·42	·31
9	—·22	—·08	·10	·00	·09	·04	·01	—·35	—·02	·12	·09
10	·25	—·09	—·28	—·03	·08	—·04	·04	·03	—·08	·17	·04
11	—·50	·41	—·94	—·07	—·57	—·06	·14	·17	·11	·09	·77
12	·09	·09	—·08	—·09	—·18	·06	—·10	·06	·02	·18	—·25
13	—·06	—·06	·12	·11	·07	·02	—·01	·00	—·17	—·05	·11
14	·17	·10	·13	—·36	—·02	·11	·05	·38	·08	·14	—·07
15	—·16	—·02	·13	—·13	·13	·06	—·09	—·59	—·06	·14	·15
16	·01	·05	·32	·02	·17	·16	·17	·17	·35	—·31	—·09
17	·09	·14	·01	—·12	·22	·05	·57	·34	·02	·10	·26
18	—·36	·00	—·62	·30	—·15	·41	—·10	·14	—·26	·20	·29
19	·18	—·07	·01	·03	·08	—·24	·01	—·03	·31	—·24	—·02
20	·01	·01	·00	—·02	—·18	·32	·03	—·15	·22	·22	·17
21	·07	·18	—·28	—·01	—·14	—·14	—·10	—·28	·01	—·18	·12
22	·13	·38	·20	—·14	—·20	—·17	·02	—·01	·08	·02	·13
23	·32	·07	·09	—·05	—·12	·45	—·05	—·02	·32	·05	·06
24	·30	·04	—·02	—·13	·01	·43	—·06	·03	·37	·35	·22
25	—·28	—·02	·08	·09	·06	—·44	·19	·03	—·35	—·23	·86
26	·06	—·06	—·12	·02	—·07	·05	—·06	·07	·33	—·09	·29
27	—·06	·00	·12	—·04	—·33	—·07	—·14	·05	·09	·10	·16

TABLE I. UNROTATED VECTORS—*continued*

Trait	Factor 1	2	3	4	5	6	7	8	9	10	11
28	·00	—·06	—·02	—·01	—·19	—·03	—·20	—·01	·09	—·14	·05
29	·06	·24	·99	·10	—·19	·04	·02	—·08	—·07	·01	—·26
30	—·22	—·04	—·29	—·01	—·18	·10	—·07	·28	—·07	—·06	·18
31	—·06	—·15	—·31	·16	—·02	·01	·02	—·08	—·13	—·01	—·10
32	—·08	·15	·06	·04	—·23	·29	·11	—·29	—·09	—·09	·14
33	—·35	—·25	—·03	·19	—·30	—·25	—·12	·03	—·04	—·08	—·19
34	·03	—·24	·52	—·07	·07	·07	·06	—·04	—·07	—·05	—·19
35	—·03	·39	·02	·27	·29	·21	—·21	·05	·05	·11	—·02
36	·00	·43	—·30	·36	·32	·01	—·30	·23	·05	—·42	—·14
37	—·43	—·05	·12	—·25	·13	—·20	—·13	—·25	—·02	·14	·34
38	—·39	—·65	—·13	—·03	—·06	—·23	·04	—·41	·07	·06	—·84
39	—·12	—·32	·35	—·19	·00	—·28	—·05	·14	—·02	—·25	·35
40	—·15	·28	·05	·50	·31	·23	—·29	·24	·06	—·12	·02
41	·08	·06	·23	·06	·20	—·06	·03	—·16	·43	·03	·04
42	—·08	·39	·21	·11	·16	·07	—·36	—·12	·01	·07	·04
43	·05	—·06	·15	—·15	—·13	·01	·09	—·02	—·15	—·33	—·08
44	—·08	·13	·06	·05	—·05	—·10	·08	·00	—·04	·09	·18
45	—·03	—·19	—·40	—·48	—·05	·06	—·05	—·15	—·03	·20	·06
46	—·17	·12	·29	—·40	·43	·07	·04	—·01	·17	—·17	—·16
47	·39	·14	·44	·29	·02	—·18	·06	·08	—·28	·17	—·80
48	·30	—·51	·10	·32	—·15	·12	—·11	·00	·20	—·04	·17
49	—·70	—·14	—·06	·20	—·18	·38	·19	—·09	·10	·14	·34
50	·04	·22	·35	—·26	—·20	·42	·16	—·29	—·37	—·64	—·22
51	·14	—·01	·39	—·31	—·27	·23	·10	—·48	—·09	—·57	—·93
52	·01	·15	—·42	·43	—·02	·04	—·28	—·26	·20	·01	·21
53	—·34	—·20	—·24	·46	·02	—·25	·24	·09	—·04	—·04	·41
54	—·17	·43	·06	·22	—·27	·25	—·04	—·09	—·01	·18	·15
55	—·06	·08	·09	·20	·35	—·05	·33	·03	—·06	·69	—·17
56	·42	·12	·08	·10	—·25	·22	·44	·04	—·48	—·35	—·46
57	—·27	—·08	—·05	—·08	—·06	—·37	·34	·19	·01	—·10	—·07
58	—·16	·16	·16	—·04	·19	·38	—·08	·00	—·31	—·29	—·38
59	—·38	·05	·30	·08	—·06	—·13	·36	·38	—·01	·17	·34
60	—·38	—·40	·12	—·03	—·10	·08	·01	·44	—·25	·13	·26
61	—·05	·15	·10	·07	—·09	·31	—·59	—·02	—·19	—·16	·06
62	·29	·11	—·39	—·04	—·16	—·11	—·04	·56	·05	·00	·23
63	·06	—·03	·26	—·04	·19	—·07	—·20	—·10	·15	—·10	—·04
64	·30	·14	·12	·00	·07	—·03	—·08	·29	—·01	—·26	—·20

III. RESULTS OF FACTORIZATION

The correlations for the above variables were calculated by the product moment formula on a 12-point scale ; and the resulting matrix was factorized by the multiple group centroid method (9). Eleven factors were extracted before inability to obtain a new group, not in the space of the existing factors, suggested that the extraction was complete. A test of the residuals confirmed this ; and the subsequent rotation showed at least 11 factors.

Rotation for simple structure proceeded with much greater difficulty than in the related study ((7) R_1F_1), and required 31 overall rotations (against 9 in R_1F_1). The percentage of variables obtained in the hyperplane (through $\pm$ ·10) was also lower (57 per cent. instead of 84 per cent.). No explanation occurs except that this factorization was done by the multiple group method which is more liable to errors in the communality and generally gives slightly larger loadings. However, this ' blindly ' obtained simple structure seemed both stable and unimprovable. The only factors where even one alternative position seemed possible were 2 and 5. In both cases the position having the worse (i.e., more oblique) angle to other factors was eventually taken, because it gave better sense when meanings were examined. Number 2 turned out to be general intelligence, which, because of its comprehensive influence, is usually the last to settle on an acceptable hyperplane.

The unrotated and rotated reference vector matrices and the correlations of the factor reference axes are set out in Tables I, II, and III respectively. The complete correlation matrix is deposited at the American Documentary Institute. In the rotated matrix the factors are arranged roughly in order of diminishing contribution to the variance of all tests.

TABLE II. FINAL ROTATED MATRIX

Trait	Factor 1	2	3	4	5	6	7	8	9	10	11	h^2
1	·03	·47	·04	—·10	·08	·36	·06	—·09	·23	—·08	·23	·49
2	—·20	—·42	·00	·02	·04	—·08	—·05	—·16	—·03	·08	—·03	·26
3	—·04	·24	·25	·01	—·19	·27	·02	—·18	·33	·19	—·02	·40
4	·01	·09	—·05	—·23	·17	—·16	·03	—·10	·15	—·41	—·02	·31
5	·02	—·07	—·11	·46	·04	·17	·00	—·07	—·03	·09	—·41	·43
6	·08	·05	—·30	—·27	·32	·31	·06	·31	—·03	—·07	·00	·49
7	·11	—·13	·55	—·01	—·37	—·34	—·06	—·04	—·35	—·09	·07	·74
8	—·04	·84	·00	—·29	·11	—·03	—·01	·08	—·07	·08	·09	·82
9	·02	·09	·35	·01	—·01	·03	·30	·00	—·05	—·03	—·08	·23
10	·30	—·12	·04	—·05	—·15	·05	—·14	—·10	·03	·09	—·10	·18
11	·57	·03	—·58	·65	·22	·02	·03	—·02	—·05	—·03	·05	—
12	·21	·10	—·05	·00	—·18	—·08	—·13	·23	—·13	—·03	—·01	·18
13	—·10	—·05	·09	·08	—·05	—·06	—·02	—·11	·07	—·10	—·17	·09
14	—·04	·23	—·22	—·05	·38	—·05	—·42	·05	—·02	·01	—·29	·38
15	·04	·04	·49	·09	—·05	—·07	·46	—·08	—·08	·06	—·25	·55
16	—·51	·23	—·26	—·27	—·12	·21	·04	·00	·16	·09	·09	·68
17	—·03	·23	—·10	·09	·33	·34	—·30	—·34	·33	·21	—·17	·56
18	·64	·06	—·06	·12	—·01	·06	—·06	—·06	·09	—·60	·09	·84
19	—·28	—·15	—·20	—·19	·11	·07	·10	·01	—·06	·32	·13	·42
20	·16	·41	·03	—·08	—·07	—·11	·15	·06	·02	—·01	—·02	·24
21	·09	—·27	—·10	·21	·07	—·08	·29	—·05	·01	·23	·03	·31
22	—·09	·06	—·08	·34	·00	—·23	—·05	—·11	—·06	·33	—·02	·32
23	·01	·40	—·17	—·23	·09	—·16	·05	·00	·13	·06	—·04	·29
24	·22	·48	—·07	—·28	·32	—·12	—·09	—·09	—·04	·06	—·21	·44
25	—·24	—·38	·10	·54	—·03	—·15	·01	—·63	·16	·06	—·04	·95
26	—·08	·08	—·29	—·16	·14	—·07	·04	—·04	—·04	·07	·05	·24
27	—·01	·13	—·08	·10	—·30	—·34	—·07	·11	—·19	—·04	·07	·20
28	—·09	—·07	—·17	—·02	·08	—·21	·06	·11	—·09	—·05	·07	·14
29	—·47	·37	·32	·17	—·04	—·38	·00	·01	·04	·07	·20	·77
30	·13	—·03	—·29	·07	·08	—·03	—·16	·06	·01	—·32	·00	·24
31	·22	—·15	·04	—·04	—·04	·14	·09	·10	·07	—·13	·12	·16
32	·02	·19	·09	·25	—·06	—·11	·34	—·04	·29	—·03	·04	·33
33	—·06	—·12	—·04	—·02	·06	—·08	·05	·41	—·22	—·25	·35	·42
34	—·35	·15	·23	—·17	—·09	—·08	·01	·08	·09	—·08	—·05	·26
35	·15	·07	·07	—·02	·01	·09	—·05	—·27	—·11	—·09	·07	·17
36	—·04	—·41	—·39	—·01	·36	·25	—·04	—·25	—·03	—·06	·23	·80
37	—·08	—·04	·29	·15	·09	—·07	·19	—·07	—·31	—·04	—·31	·38
38	·00	—·07	·31	—·46	·04	·38	·40	·95	—·29	—·16	·20	—
39	—·51	—·15	—·08	—·01	·22	—·25	—·09	—·10	—·06	—·04	—·14	·45
40	—·03	—·02	—·10	—·14	·21	·10	—·09	—·30	—·08	—·30	·27	·38
41	—·23	·18	·05	—·24	·27	·09	·18	—·07	—·15	·32	·07	·40
42	·01	·00	·16	·12	—·01	—·16	—·06	—·22	—·24	—·04	—·03	·22
43	—·29	—·10	—·10	·12	·02	—·07	·08	·07	·26	·00	—·05	·22
44	·03	·04	·07	·20	—·07	—·05	—·02	—·13	—·02	·07	·04	·08
45	·31	—·03	·01	—·02	—·06	·03	·05	·21	—·09	—·04	—·46	·37
46	—·43	·08	—·01	—·15	·28	·30	·09	·00	—·05	·08	—·41	·56
47	·03	—·05	·34	·00	—·58	—·02	—·30	·13	·00	·13	·34	·70
48	·09	—·10	·05	·44	—·12	·29	—·02	·01	—·06	·09	—·30	·40
49	·09	·50	·09	—·01	·16	·14	·29	·15	·07	—·45	·16	·64
50	—·45	·00	—·04	·38	—·06	—·09	·43	·03	·69	—·07	—·18	—
51	—·51	·02	·02	—·02	—·03	·05	·56	·65	·39	·10	·00	—
52	·33	—·13	—·05	—·03	·22	—·03	·30	—·14	—·15	·05	·31	·42
53	·06	—·14	·01	·06	·18	·19	·05	—·19	·05	—·09	·39	·30
54	·23	·32	·07	·33	—·16	—·17	·10	—·27	·01	—·09	·19	·69
55	·41	·31	·55	—·08	—·44	·37	—·24	—·12	—·14	·08	·04	—

TABLE II. FINAL ROTATED MATRIX—*continued*

Trait	Factor 1	2	3	4	5	6	7	8	9	10	11	h^2
56	—·05	—·07	—·05	·28	—·42	·07	—·02	·03	·77	·06	·19	·98
57	—·18	—·08	—·14	·11	·11	·27	—·09	·19	—·01	·10	·10	·23
58	—·16	—·04	·06	·04	—·06	·16	·09	·03	·26	—·36	—·14	·30
59	·16	—·31	—·01	—·20	·04	—·05	·30	·14	·00	·09	—·11	·31
60	—·03	·14	·00	—·05	—·11	—·14	—·39	·04	—·02	—·58	—·08	·55
61	·01	—·11	—·04	·09	—·01	—·43	·03	—·10	—·08	—·42	—·06	·41
62	·21	—·15	—·54	·07	·02	—·11	—·53	—·15	—·01	·05	·01	·68
63	—·27	—·06	·06	—·17	·17	—·08	·09	—·05	—·16	·09	—·06	·20
64	—·22	—·17	—·27	—·05	·04	—·04	—·25	—·09	·09	·07	·04	·24
Number in ± ·10	·32	·30	·39	·32	·32	·31	·41	·36	·39	·46	·36	—

The communalities (h^2) are those of the *unrotated* matrix. Occasional poor estimates of communality in the original calculations have resulted in these few slightly exceeding unity. These are left blank.

TABLE III. CORRELATIONS OF FACTOR REFERENCE AXES

Factors	1	2	3	4	5	6	7	8	9	10	11
1											
2	·04										
3	·31	·05									
4	·16	—·24	·12								
5	—·59	—·07	—·58	—·43							
6	·00	·05	—·03	—·25	·30						
7	—·19	·01	·21	—·03	·37	·18					
8	—·02	·27	—·22	—·29	·12	·08	—·14				
9	—·13	·06	—·06	·21	—·12	·21	·21	—·30			
10	—·13	—·10	—·01	·19	·11	·15	·24	·19	·03		
11	—·03	·08	—·12	—·07	·10	·00	·12	·18	·00	·05	

The factors obtained must, for reasons of space, be set out briefly with only a few lines of explanation. A full discussion will be given elsewhere (10) when evidence from several other factor researches can be simultaneously put forward. The highest 8 to 12 variables are set out below, using the same index system as that given in the Appendix, and reversing the verbal definitions for variables with negative loadings, so that the descriptions are consistent.

Factor 1

Index Number	Matrix Number	Loading	Title Variable (Appropriate to Sign of Loading)
38	(18)	·64	High Ratio of Consonant to Dissonant Statements Recalled
21	(11)	·57	Absence of Questionable Preferences
36	(16)	—·51	Inability to State Logical Assumptions
75	(51)	—·51	Little Metabolic Rate Change with Stimulation
64	(39)	—·51	Female Sex of Student
55	(29)	—·47	Low Ratio Emotional to Non-Emotional Associations
74	(50)	—·45	Low Basal Metabolic Rate
70	(46)	—·43	Rapid Appearance of Negative After-Image
57	(31)	·22	Sympathetic Increase in Emotional Recall
60b	(63)	—·27	Little Influence of Frustration on Unrelated Recall
14	(10)	·30	Many Jokes Considered Funny

This contains the two markers, 21 and 36, put in to locate the old (R_1F_1) factor 3, and is in other ways consistent with that factor. From its general character it was described as "feminine, unforceful, dependent"; and in the present research this is confirmed by the inclusion of variable 64—the sex of the student—which has its highest loading here. This is almost certainly the sex-difference factor in objective tests.

Factor 2

Index Number	Matrix Number	Loading	Title Variable (Appropriate to Sign of Loading)
11	(8)	·84	High Verbal and Numerical Ability, A.C.E. Test
73	(49)	·50	Healthy Schneider Index
50	(24)	·48	High Alertness in Following Directions
1	(1)	·47	High Fluency of Association
2	(2)	—·42	Low Perseveration (disposition rigidity)
41	(20)	·41	High Two-Hand Co-ordination
49	(23)	·40	High Ratio of High Level to Low Level Mental Sets

This picks up the markers 11 and 1 inserted to locate factor 6 of the original study, and is in other ways consistent with the interpretation that it represents general intelligence. Variable 73 offers a puzzle ; but the high position of the A.C.E. test, No. 11, and other considerations leave no doubt of the identification proposed. It may be, however, that this factor should be split in two, since it contains the only appreciable loadings on the old factor 10 (low perseveration-rigidity and low sway suggestibility).

Factor 3

Index Number	Matrix Number	Loading	Title Variable (Appropriate to Sign of Loading)
21	(11)	—·58	Questionable Preferences in Reading
9	(7)	·55	Slow Tempo
79	(55)	·55	High Eidetic Imagery
84	(62)	—·54	Lower Increase in Pulse Rate Under Stress
35	(15)	·49	High Suggestibility to Authority
71	(47)	·34	Erroneous Reactions when Maintaining Complex Mental Sets
13	(9)	·34	High Oscillation
85	(64)	—·27	Small Sympathetic Increase in Timidity Associations

This is the old factor No. 1, as shown by the markers 9 and 13. Though 9 also occurs in another factor, it there has a much lower loading. The original tentative interpretation of factor 1 was that it is either cyclothymia or surgency. 21, 84, and 13 are compatible with cyclothymic lack of tension.

Factor 4

Index Number	Matrix Number	Loading	Title Variable (Appropriate to Sign of Loading)
21	(11)	·65	Absence of Questionable Preferences in Reading
51	(25)	·54	High Index of Carefulness
5	(5)	·46	High Ratio of Regularly Warned to Irregularly Warned Reaction Times
72	(48)	·44	Slowing of Reaction Time by Complex Instructions
74	(50)	·38	High Basal Metabolic Rate
43	(22)	·34	Large P.G.R. Mean Deflection
78	(54)	·33	Lack of Muscle Tonus
29b	(44)	·20	Large Number of Extra Punitive Criticisms of Test

Variables 21 and 43 were the two highest in R_1F_1 factor 2 ; and, though 21 occurs in two other factors, because of its high communality, it is nevertheless highest here. The original factor 2 was identified with Thurstone's perceptual factor *F* and with some sort of will-character. The latter is strengthened by the new associates. The total collection of variables now includes carefulness and exactness, good preferences in reading, slowing of reaction time in adjustment to complex instructions, high aspiration level, prolonged maintenance of ergograph performance level, high level in new task from the start, large P.G.R. deflection (associated (5) with 'desire to excel'), and slow speed of closure and dark adaptation (16) : in the light of their associations we may suppose that the last two traits are partly influenced by an exacting personal standard, which does not permit *S* to react until he is absolutely sure. Thus the factor is one of perseverance, determination, aspiration, and self-control. Factor *K* (5) is the best hypothesis at present for this pattern.

Factor 5

Index Number	Matrix Number	Loading	Title Variable (Appropriate to Sign of Loading)
71	(47)	—·58	Few Erroneous Reactions when Maintaining Complex Mental Sets
79	(55)	—·44	Poor Eidetic Imagery
80	(56)	—·42	P.G.R., Little Upward Drift
36	(16)	·38	Ability to State Logical Assumptions
9a	(7)	—·37	Fast Tempo
61	(36)	·36	High Influence of Frustration of Judgement
39	(19)	·33	High Ratio of Colour to Form in Sorting
8	(6)	·32	High Rate of Cube Fluctuation

This factor has only one of the markers for R_1F_1 factor 4, namely, 8 ; but its loading here is higher than elsewhere, and the general sense of the factor is the same as the old No. 4, which was called ' general psychomotor efficiency,' and contingently identified with Thurstone's perceptual factor *D*. To match with 4 it would be necessary to consider ' fast tempo ' equivalent to ' ideomotor speed ' and variable 71 above equivalent to ' freedom from false reaction-time reactions.' The missing marker is 13, which should have a negative loading, as ' freedom from oscillation,' but actually has zero loading. This pair has the correct sign pattern, but at a lower level, in the present 3.

Factor 6

Index Number	Matrix Number	Loading	Title Variable (Appropriate to Sign of Loading)
83	(61)	—·43	Fidgetometer Frequency
55	(29)	—·38	Low Ratio Emotional to Non-Emotional Associations
79	(55)	—·37	Good Eidetic Imagery
1	(1)	·36	High Fluency
37	(17)	·34	Ability to Suggest Classifications
9	(7)	—·34	Fast Tempo
54	(28)	—·21	Planning Ability

This also is a new factor ; but only variable 54 is unique to it. Fidgetometer frequency is split between this and factor 8. This factor suggests Jaensch's *B*-type or ' surgency ' (5). It has the right pattern for the old factor 9, but at too low a level, and the present 11 is therefore better matched with the old 9.

Factor 7

Index Number	Matrix Number	Loading	Title Variable (Appropriate to Sign of Loading)
75	(51)	·56	Much Metabolic Rate Change with Stimulation
84	(62)	—·53	Little Increase of Pulse Rate in Stress
35	(15)	·46	High Suggestibility to Authority
74	(50)	·43	High Basal Metabolic Rate
34	(14)	—·42	Immaturity of Opinion
82	(60)	—·39	Low Speed on Finger Maze
25b	(32)	·34	High Ratio Emotional to Non-Emotional Recall

This carries the markers—34 and 35—from the old factor 8. It is also co-operative (5) with the present factor 1 (by variables 74 and 75), suggesting some physiological second-order factor between them. Otherwise there is no clear indication of its nature, except perhaps of some temperamental basis.

Factor 8

Index Number	Matrix Number	Loading	Title Variable (Appropriate to Sign of Loading)
75	(51)	·65	High Metabolic Rate Change with Stimulation
51	(25)	—·63	Low Index of Carefulness
37	(17)	—·34	Inability to Suggest Classifications
65	(40)	—·30	High Logical Consistency of Attitudes
58	(33)	·41	Much Myokinesis Drift
60	(35)	—·27	Low Influence of Unrelated Frustration on Recall
8	(6)	·31	High Rate Cube Fluctuation
22	(12)	·23	High Ratio Purposeful to Chance Observation
24b	(57)	·19	Much (late) Improvement in Mirror Drawing

In the old series it has a faint resemblance to factor 4, to be discounted because 4 is already matched. Old factor 10, which is unmatched, has some sign loadings on the markers 2 and 42; and these appear (apart from the intelligence factor) in the present 7, 8, and 10. Since 7 and 10 are matched the present factor 8 may be accepted as a *probable* match for the former 10.

Factor 9

Index Number	Matrix Number	Loading	Title Variable (Appropriate to Sign of Loading)
80	(56)	·77	P.G.R. Upward Drift
74	(50)	·69	High Basal Metabolic Rate
75	(51)	·39	High Metabolic Rate Change with Stimulation
62	(37)	—·31	Low Ratio of Emotional Interesting to Non-Emotional Dry Reading Speed
9	(7)	—·35	Fast Tempo
3	(3)	·33	Speed of Judgement
68	(43)	·26	Large Number of Personal Excuses
67	(42)	—·24	Few Immoderate Statements (Non-Extreme Viewpoint)
25b	(32)	·29	High Ratio Emotional to Non-Emotional Recall
22	(12)	—·13	Low Ratio Purposeful to Chance Memory

Variable 3 is the highest in the old factor 11 and has its highest loading in the present study here. Variable 22 from the old 11 has the correct sign but only a low loading —13. It should be noted that the present factor 8 is the only other competitor for the old 11, having loadings of −18 and + 23 in these variables. The variables here indicate an over-stimulated, perhaps insecure person. The large P.G.R. upward drift indicates usually an initial excessive preparedness followed by relapse into fatigue. The factor has a marked resemblance to the sympathetic reaction pattern (5).

Factor minus 10

Index Number	Matrix Number	Loading	Title Variable (Appropriate to Sign of Loading)
38	(18)	·60	High Ratio Consonant to Dissonant Opinions Recalled
82	(60)	·58	High Speed on Finger Maze
73	(49)	·45	Healthy Schneider Index
83	(61)	·42	High Fidgetometer Frequency
4	(4)	·41	Long Reaction Time
81	(58)	·36	High Regard for Instructions
43	(22)	—·33	Small P.G.R. Mean Deflection
39	(19)	—·32	Low Ratio of Colour to Form in Sorting
56	(30)	·32	Sympathetic Increase in Emotional Association

This has the two markers—38 and 39—not appearing in such loadings elsewhere, for the R_1F_1 factor 5. The old factor 5 had been considered a 'rigidity–low energy' pattern, possibly schizoid factor, $H-$ or $C-$. The present picture is consistent, adding a kind of timid, feverish following of instructions (81, 82, 83) to a small basic reactivity (4, 43, 73).

Factor 11

Index Number	Matrix Number	Loading	Title Variable (Appropriate to Sign of Loading)
69	(45)	—·46	Tendency to Disagree with Authority and Sample
70	(46)	—·41	Rapid Appearance of Negative after Image
5	(5)	—·41	Low Ratio of Regular to Irregular Reaction Time
77	(53)	·39	High Salivary pH (Alkaline)
58	(33)	·35	High Upward Myokinesis Drift
71	(47)	·34	Many Erroneous Reactions when Maintaining Complex Mental Sets
76	(52)	·31	High Pitch of Voice
62	(37)	—·31	Low Ratio of Emotional Interesting to Non-Emotional Dry Reading Speed
27	(13)	—·17	Low Impairment of Memory by Emotion

By exclusion, and because of the high loading in 5, this *might* be R_1F_1 factor 9, which Thurstone's perceptual factor *C* had been tentatively matched. The other marker, 4, has no loading. Only the present 4 and 6 have these markers in correct pattern and 4 is matched. It has several somatic variables, sometimes associated with absence of fatigue, but is of too low a variance to permit any definite hypothesis.

IV. SUMMARY

1. A multiple factor analysis of 42 new objective personality tests, with 22 older tests inserted to give continuity and mark factors from an earlier study (7), has resulted in 11 factors, which were rotated to a tolerable simple structure.

2. All but one, or possibly two, of the 11 factors of the older study can be recognized by their markers here. In most cases the two intended markers stand out clearly ; but for the factors numbered 5 and 11 only one marker appears.

3. There thus seems to be one entirely new factor (or possibly two) in the present study, numbered 6 (and 8), and one (or possibly two) in the last, numbered 7 (and 10), which have found no match here. The high variables in 6 (and 8) are all new. The absence of 7 is inexplicable ; but 10 had such small variance before that its match with 8 here is necessarily tentative.

4. The new factor 6 has the general nature of energy, confidence, speed, and logical planning ability (suggesting rating factor *I* or *G*). Factor 8 develops the old, ill-defined 10 in the direction of carelessness, physiological over-reactivity, and mental confession, which is possibly consistent with the interpretation of the old 10 as general neuroticism (13).

5. The seven adequately confirmed factors (and, to a less extent, the three less clearly confirmed) are each enriched by new variables in the pattern, which will permit clearer hypotheses to be made as to their psychological nature. These hypotheses are discussed elsewhere in setting up further tests for the purpose of obtaining more saturated factor measures (10).

This investigation was supported (in part) by a research grant from The National Institute of Mental Health, of the National Institutes of Mental Health, Public Health Service, Washington, U.S.A.

APPENDIX

Description of Tests

The first 22 tests are listed only by titles, their description being available in the earlier study (7).

Master Index No.	Matrix No.	Reliability	Title and Description
1	(1)	·68	*Fluency of Association* (In (7) No. 1.) Marker for R_1F_1 factor 6.
2	(2)	·41	*Perseveration or Disposition Rigidity.* (In (7) No. 2.) Marker for R_1F_1 factor 10.
3	(3)	—	*Speed of Judgement.* (In (7) No. 3.) Marker for R_1F_1 factor 11.
4	(4)	·73	*Slowness of Reaction Time.* (In (7) No. 4.) Marker for R_1F_1 factor 9.
5	(5)	—	*Ratio of Regularly Warned to Irregularly Warned Reaction Times.* (In (7) No. 5.) Marker for R_1F_1 factor 9.
8	(6)	·68	*Alternating Perspective* (Cube Fluctuation). (In (7) No. 8.) Marker for R_1F_1 factor 4.
9	(7)	—	*Tempo.* (In (7) No. 9.) Marker for R_1F_1 factor 1.
11	(8)	·84	*Verbal and Numerical Ability on the A.C.E. Test.* (In (7) No. 11.) Marker for R_1F_1 factor 6.
13	(9)	·44	*Oscillation.* (In (7) No. 13.) Marker for R_1F_1 factor 4 (and 1).
14	(10)	·76	*Number of Jokes Considered Funny.* (In (7) No. 14.) Marker for R_1F_1 factor 1.
21	(11)	·28	*Absence of Questionable Preferences in Reading.* (In (7) No. 21.) Marker for R_1F_1 factor 2 (and 3).
22	(12)	—	*Ratio Purposeful to Chance Observation and Memory.* (In (7) No. 22.) Marker for R_1F_1 factor 7 (and 11).
27	(13)	·32	*Impairment of Mental Performance by Embarrassing Emotion.* (In (7) No. 27.) Marker for R_1F_1 factor 8.
34	(14)	—	*Immaturity of Opinion.* (In (7) No. 34.) Marker for R_1F_1 factor 8.
35	(15)	—	*Suggestibility to Authority.* (In (7) No. 35.) Marker for R_1F_1 factor 8.
36	(16)	·73	*Ability to State Logical Assumptions.* (In (7) No. 36.) Marker for R_1F_1 factor 3.
37	(17)	·88	*Ability to Suggest Classifications.* (In (7) No. 37.) Marker for R_1F_1 factor 3.
38	(18)	·49	*Ratio Consonant to Dissonant Statements Recalled.* (In (7) No. 38.) Marker for R_1F_1 factor 5.
39	(19)	—	*Colour-Form Ratio in Sorting.* (In (7) No. 39.) Marker for R_1F_1 factor 5.
41	(20)	—	*Two Hand Co-ordination.* (In (7) No. 41.) Marker for R_1F_1 factor 7.
42	(21)	·87	*Body-Sway Suggestibility.* (In (7) No. 42.) Marker for R_1F_1 factor 10.
43	(22)	—	*Magnitude of Mean P.G.R. Deflection.* (In (7) No. 43.) Marker for R_1F_1 factor 2.
49	(23)	·71	*Ratio of Higher to Lower Mental Set Performance.* Task requires maintenance of set at three levels : (1) to read every other word in a passage (arranged to make sense) ; (2) to note for reproduction all objects mentioned in certain categories (colours, personal characteristics) ; (3) to note objects *simultaneously* falling in three of these categories. *S* is scored on ratio of success in highest to other two tasks. Theory is that schizoid ($A-$ factor) person will do less well on higher sets.
50	(24)	·72	*Alertness in Following Directions.* *S* is here scored on *sum* of successful observations (recalled immediately after the performance) in (2) and (3) above. The hypothesis is that high *G* factor (5) will be associated with care and conscientiousness in this alertness-demanding situation.
51	(25)	·49	*Index of Carefulness.* Also derived from 49. *S* is scored in middle and upper sets on the number of instances wrongly categorized. Row score is inverted to give ' carefulness ' not errors. Assumed to express both $G-$ and $F-$ factor (Desurgency).
52	(26)	—·45	*Tendency to Remember Easier Tasks.* *S* is asked to solve two pages of riddles, those on one page being more difficult than the next. Score is number recalled after difficult page *minus* number recalled after easy page. Presumed to test dominance factor, $E-$.
53	(27)	·01	*Ratio of Self-Criticism to Self-Approval.* *S* writes down in one minute single words or short phrases describing his undesirable and desirable characteristics. The ratio is number of self-criticisms to number of self-approvals. Presumed to test factors $A-$ or $E+$.

Master Index No.	Matrix No.	Reliability	Title and Description

54 (28) ·58 *Lack of Planning Ability.* Two tasks used : making words from groups of letters and picking, from nine numbers, pairs that would sum to a given number. The composite score was based on how evenly the *S* distributed his efforts, how much he used scratch paper, and how systematic he was in arithmetical section. Deliberation and planning conduces to a low score. Possibly a test of factors *C*, *G*, or *I*.

55 (29) ·63 *Ratio of Emotional to Non-Emotional Association. S* given a key word and two choices of associative words. In each, one word was more emotional than the other. Score is the ratio of emotional to non-emotional association words. Assumed to test emotionality of *C*— factor.

56 (30) ·23 *Sympathetic Increase in Emotional Associations.* Test 55 was in two parts with a story of fear interposed. Score, the number of emotional words chosen on the second page *minus* the number of emotional words chosen on the first page. Assumed to test *A* + or *C*— factor.

57 (31) ·21 *Sympathetic Increase in Emotional Recall.* A list of words, some emotional and some matter of fact, was presented ; *S* then recalled them three times, (i) directly after presentation, (ii) after reading of a ' fearful ' story, and (iii) after reading of several jokes. The increase in number of emotional words recalled was measured.

87 (32) ·01 *Ratio Emotional to Non-Emotional Recall.* In the preceding test the ratio of emotional to non-emotional words recalled was taken for the first situation (before sympathetic effect). Assumed to test *H* factor.

58 (33) ·44 *Myokinesis Drift. S*s were asked to draw two series of ten vertical lines under Mira's conditions (see (5)). The distance of the base line of these lines from a horizontal line through the starting point at a point 6 ins. along the base line is the measure of drift ; upward drift is positive. Assumed to test *C*— (neuroticism).

59 (34) ·07 *Effect of Tempting Distraction upon Recall.* A scientific passage was read by *S*, and he was tested on recall of what he had read. After reading to the middle of an interesting story *S* was again tested for knowledge of the passage. Score number of correct answers to first test *minus* number of correct answers to second. A negative or low positive score indicates no distraction. Assumed to measure poor personality integration, i.e., low *G* factor.

60 (35) ·28 *Influence of Unrelated Frustration on Recall.* Instead of finishing the interesting story *S* was given another task and then questioned on the original passage. This interruption was quite frustrating. The score was the number of correct answers to the first test minus the number of correct answers to this. Intended as a measure of *C* factor.

61 (36) —·13 *Influence of Frustration on Judgement.* Two separate pages of judgement tests (underlining the word that did not belong). The second page followed the frustrating situation in 60. Score for correct answers on the first page *minus* the number on the second page (after frustration). Intended as a measure of the emotionality of *C* and *D* factors, mostly the latter.

62 (37) — *Ratio of Emotionally Interesting to Non-Emotional* (*Dry*) *Reading Speed.* The number of words read in an interesting story was divided by the number read in the same length of time in a scientific passage. It was assumed that a high *G* individual would maintain good speed in the uninteresting material.

63 (38) —·05 *Masculinity versus Femininity* (Terman-Miles). Questions about preferences in (1) subject matter for drawing and reporting, (2) things to do on vacations, (3) famous people. Terman and Miles' weighted scores were used (15). Inserted to explore possibility that *I* factor may be masculinity.

64 (39) — *Sex of Student* (Positive for Male).

65 (40) ·23 *Poor Logical Consistency.* In this test the subject expressed his attitudes on a wide range of issues. The attitude or belief statements dispersed through the test could be gathered together in groups of three, deliberately designed to present the logical coherence of a first and second premise and conclusion in a syllogism. *S* was scored on the number of consistent syllogisms. It is assumed that the low integration of *C*— factor will show itself in an occasional need to accept logically inconsistent views.

66 (41) ·00 *Imaginative Synasthesia.* Modernistic drawings were presented to *S* along with a series of abstract terms, e.g., coyness, boredom, doubt, etc., which the artist (Abner Dean) had used as titles for his drawings. *S* was asked to choose the best title. It is assumed that this artistic sensitivity is the essence of *I* factor.

67 (42) ·68 *Immoderation of Statement.* In tests 34, 35, and 65 and 69 *S* was scored for the number of extreme answers (' absolutely certain ' and ' absurd or stupid ').

68 (43) ·26 *Personal Excuses.* At the end of the testing week the subject was asked to list ways in which various tests and ratings might be ' unfair ' to him. Score : number of excuses.

88 (44) ·13 *Extra-Punitive Criticisms. S* was asked for criticisms of the tests in general. Score : the number of remarks blaming examiner or external circumstances for anticipated poor score.

Master Index No.	Matrix No.	Reliability	Title and Description

69 (45) ·78 *Tendency to Agree with Majority.* Statements by authorities were given and below each were presented the percentage of students agreeing and disagreeing. Score: number of times *S* agreed with higher percentage. (Overlaps with 35 above.)

70 (46) ·90 *After-Image Time.* The length of time between the disappearance of the positive after-image and the appearance of the negative after-image (indicated by *S* pressing a key). Four trials in dark box after standard exposure. Assumed that in neurasthenia after-images are unduly long, and perhaps affected by desurgency.

71 (47) ·32 *Erroneous Reactions when Maintaining Complex Mental Sets.* *S* reacts to a complex set of lights and buzzers by pressing a key with his left or right hand, or not at all, according to the sequence. Score : number of errors. Assumed to measure high excitability and low frustration tolerance.

72 (48) ·67 *Slowing of Reaction Time by Complex Instructions.* The reaction time for correct reaction in the above complex task was divided by the *S*'s normal, simple reaction time to visual and auditory stimuli. Possibly an expression of the general level of inhibition and therefore of *F* factor.

73 (49) ·86 *Healthy Schneider Index.* The usual composite index of changes in pulse rate and blood pressure from lying down to standing and time to return to normal after exercise. Added for evidence on psychosomatic theories.

74 (50) ·99 *Basal Metabolic Rate.* The *S*'s smallest O_2 consumption for four runs (six minutes each), converted to calories/hr./sq. metre of body area by reference to sex-age tables. Introduced to test relations with *H* factor (see 14).

75 (51) ·51 *Metabolic Rate Change with Stimulation.* Measured under three types of stimulation : unusual arithmetic problems, muscular work (ergograph), and emotional stimulation (electric shock and loud noise). Intended as a measure of sympathetic reactivity (hypothesized as *D* or *H* factors).

76 (52) — *Pitch of Voice.* While *S* was answering questions his voice was matched for essential pitch with one of nine grades on a standard voice record.

77 (53) — *Salivary pH.* *S* accumulated saliva for one minute. pH was read one minute after ejection on the usual electrical pH meter.

78 (54) ·63 *Lack of Muscle Tonus.* While *S* was reading, his left arm rests on a trolley. The arm was suddenly snatched backwards by a falling weight attached to the trolley and the distance the arm moved was recorded by a stylus tracing interrupted after $\frac{1}{5}$ second. Designed to measure resting muscle tonus, as representing a temperament factor, probably *J*.

79 (55) ·55 *Eidetic Imagery.* Following Allport's findings, eidetic imagery proneness is measured with eight pictures chosen for subtlety of colour and form. Points given for externalizing and locating the picture, sensations of eye accommodation, etc., colour *minus* pure form, non-verbalizable form *minus* verbalizable form. The subtractions ruled out goodness of memory. Hypothesis : that Jaensch's types may be unitary personality factors already known.

80 (56) — *P.G.R. Upward Drift.* After a task in which the *S* was shocked, shock apparatus was disconnected and *S* was assured he could relax. The P.G.R. upward drift for three minutes was recorded. Hypothesis : measures cyclothyme's trustful readiness to relax (*A* + or *L*— factor).

86 (57) ·44 *Improvement (Learning) in Mirror Tracing under Stimulation.* Three trials were given *S* on each of three patterns. In the second the *S*s were shocked for poor performance. Shock was also administered on the first pattern of the third trial. The second was done without mirror, as a control measure of dexterity. Score was reduction of error score between first and third performance. Hypothesis : that sheer mirror drawing ability is cancelled by the comparison and this represents effective performance under stress or strength of escape motivation.

81 (58) ·48 *Carelessness over Following Instructions.* Number of slips made by *S* on three mirror tracings and three ordinary tracings of designs. Hypothesis : that carefulness measures *G* factor.

89 (59) ·63 *Impairment of Performance by Fright.* Three equally difficult finger mazes, the first done under normal conditions, the last two under conditions of fright (electric shocks on maze 2 and a rat and some cockroaches in the cage surrounding the maze 3). Score : the turns per minute for maze 1 divided by the turns per minute in mazes 2 and 3. Intended to test the emotionality factors.

82 (60) ·59 *Speed on Finger Maze.* Number of turns per minute on five finger mazes. Intended to measure readiness for trial-and-error in *H* factor.

83 (61) ·53 *Fidgetometer Frequency.* A pneumatic instrument shows number of body movements per minute on a moving drum, especially of leg and foot. Score : movements per minute during test 79. Hypothesis : muscle tension in a standard situation more sensitive than P.G.R., especially for tension of schizothyme.

84 (62) — *Increase of Pulse Rate in Stress.* Difference in normal P.R. and the P.R. after the frightening mirror tracing shock-situation was recorded. Presumed to measure *D* or *I* factor.

Master Index No.	Matrix No.	Reliability	Title and Description
90	(63)	·38	*Loss of Recall from Frustration and Humiliation.* *S* memorizes two lists of phrases and recalls each after interpolated activity. Score : difference in number of phrases recalled after a neutral experience (easy riddles) and after a frustrating or humiliating task. Hypothesis : an inverted measure of ego strength, i.e., of *C* factor.
85	(64)	—	*Sympathetic Increase in Timidity Associations.* Similar to test 5 except that only the increase in associative words appropriate to a fear emotion was counted. Hypothesis : a measure of *H*— factor, indicating proneness to timid withdrawal responses.

Note.—Variables 52, 53, 59, 61, 63, 66, and 87 are retained in the matrix for hyperplane but not involved in any definition or discussion of factors beyond this point because of the negligible reliability which they proved to have.

REFERENCES

1. Cattell, R. B. (1932). ' Temperament tests : I. Temperament.' *Brit. J. Psych.*, XXIII, 308–29.
2. Cattell, R. B. (1933). ' Temperament tests : II. Tests.' *Brit. J. Psych.*, XXIV, 20–49.
3. Cattell, R. B. (1943). ' Fluctuation of sentiments and attitudes as a measure of general character integration and temperament.' *Amer. J. Psychol.*, LVI, 195–216.
4. Cattell, R. B. (1945). ' The description of personality. III. Principles and findings in a factor analysis.' *Amer. J. Psychol.*, LVIII, 69–90.
5. Cattell, R. B. (1946). *The Description and Measurement of Personality.* New York : World Book Co.
6. Cattell, R. B. (1947). ' Confirmation and clarification of primary personality traits.' *Psychometrika*, XII, 197–220.
7. Cattell, R. B. (1948). ' Primary personality factors in the realm of objective tests.' *J. Person.*, XVI, 459–87.
8. Cattell, R. B. (1950). ' The main personality factors in questionnaire, self-estimate material.' *J. Soc. Psychol.*, XXXI, 3–38.
9. Cattell, R. B. (1951). *Factor Analysis in Psychology.* New York : Harper Bros.
10. Cattell, R. B. ' Review of interpretations of the primary personality factors.' (In press.)
11. Cattell, R. B., Dubin, S. S., and Saunders, D. R. ' Personality structure in psychotics as determined by factorization of objectively measured behaviour responses.' (In press.)
12. Cattell, R. B., and Saunders, D. R. ' Verification of factors in one hundred and six objective personality test designs.' (In press.)
13. Eysenck, H. I. (1947). *Dimensions of Personality.* London : Kegan Paul.
14. Herrington, L. P. (1942). ' The relation of physiological and social indices of activity level.' In McNemar, Q., and Merrill, M. A. *Studies in Personality* (in honour of L. M. Terman). Stanford Univ. Cal. Stan. Univ. Press.
15. Terman, L. M., *et al.* (1936). *Sex and Personality.* New York : McGraw Hill.
16. Thurstone, L. L. (1944). *A Factorial Study of Perception.* Chicago : Univ. of Chicago Press.

Vol. IV
Part III

The British Journal of Psychology
Statistical Section

November,
1951

FACTOR ANALYSIS OF ASSESSMENTS FOR A SINGLE PERSON

By CYRIL BURT and HARRIET WATSON
University College, London

I. *The Correlation of Persons.* II. *Illustrative Case:* A. *The Case History.* B. *Correlations and Factor Analysis.* C. *The Factor-Measurements.* III. *Summary.*

I. THE CORRELATION OF PERSONS

Critics of factor analysis frequently complain that it reduces the individual personality to a sum of unrelated assessments for a series of isolated traits, and thus affords no adequate picture of the concrete pattern or configuration formed by the relations between component traits in one and the same person. Such criticisms may no doubt seem plausible so long as we have in mind a factorization of the commoner type of correlation matrix, namely, a table of correlations between tests or traits. But, as has been pointed out in previous papers (3, 4, 6, and refs.), it is often desirable, and at times even essential, to correlate not tests but persons. Now when we factorize a table of correlations in which the variables are persons or aspects of a person, and in which the sample consists not of selected individuals but of selected traits, then the factors obtained, together with their schemes of numerical assessments, will evidently indicate the kind of patterns which those traits produce. Each separate set of factor-measurements will specify a unique configuration of characteristics, describing, not the composition of particular tests, but the constitution of particular persons or types of person. This alternative mode of statistical approach has been so constantly overlooked that it seems desirable to describe in some detail an actual example, which will demonstrate the results of the procedure and illustrate the method of calculation.

Our contention can perhaps be most clearly exemplified when the assessments analysed relate to a single person only.[1]

The extent to which the possibilities of this kind of investigation have escaped notice is shown in the long-standing dispute between the champions of clinical psychology and of statistical psychology —the ' ancient battle between intuitionist and psychometrist,' as Spearman has described it. The clinical psychologist stresses the importance of a systematic study of the single personality and " the insight which " (as he contends) " can be gained into human nature by this method and no other " : his conclusion usually is that " the correlation or factorization of data from a *mass* of individuals can yield no information about any *single* individual, since each single individual is unique." The psychometrist retorts that " the inferences drawn from a single case can be no more than unconfirmed impressions," and maintains that " verifiable laws can only be reached by statistically analysing *groups* of cases." Curiously enough, neither side appears to realize that statistical analysis is quite as applicable to the " study of the single personality " as to the study of groups, so

[1] This procedure has been in occasional use for many years. The earliest instance, so far as we are aware, was an attempt by Burt to study the changes in the relative preferences for different subjects of the curriculum shown by one and the same child with changes of teacher, school, and age (*Report of L.C.C. Psychologist*, 1927, pp. 13–21). Here it was found that the study of the preferences shown by *groups* of children at different ages tended to obscure the striking contrasts between those particular *individuals* whose preferences remained pretty constant and those whose preferences were easily swayed by altered circumstances. Miss Davies' review of investigations in which the method of correlating persons has been used (10) includes very few published articles in which only a single person has been studied in this way. But more recently the method has been adopted to supplement other techniques in several unpublished theses. (On the method itself see Burt, C., ' Correlations between Persons,' *Brit. J. Psych.*, XXVIII, 1937, pp. 59f.)

that the " impressions gained " in the former case can after all be subjected to rigorous verification by means of available statistical techniques (see 12 and refs.).

Contemporary critics are more cautious. They admit that cognitive qualities can be studied by statistical procedures ; but deny that these are applicable to the orectic or ' dynamic ' aspects. A cognitive test, so Miss Ainsworth tells us, " deals with one variable or one function," and so can reasonably claim to " place all tested individuals on a single continuum with respect to that function " ; on the other hand, when the psychologist turns from the cognitive to the orectic aspect, he has to deal with " *n* functions or variables, and to describe the individual in terms of a dynamic pattern of interrelated functions or variables " (cf. 1, p. 151). We ourselves do not altogether agree with the contrast as thus stated, though it may seem borne out by the older and simpler uni-factor methods of Spearman and his followers. But these critics have apparently quite forgotten that, with multi-factor procedures, it is perfectly possible to obtain factors which will deal with " *n* functions or variables," and express " patterns." Further, since (to a first approximation) the factors are treated as combining after the analogy of physical forces, they may reasonably be described as ' dynamic ' (if indeed we feel that the use of this term in psychology is really helpful).

II. ILLUSTRATIVE CASE

As part of an investigation into the influence of puberty on the emotional characteristics of children in London County Council schools an attempt was made, with the assistance of teachers and research students, to obtain repeated assessments for the same individual children, at equidistant intervals of time, over a period of several years, extending from about 12 or 18 months before the advent of puberty to about 12 or 18 months afterwards.[1] To ascertain which components in the child's personality remained stable or invariant throughout the period, and which were liable to be disturbed by internal or external influences, the most natural procedure is to subject the data obtained in each case to a factorial analysis. Since a single observer might see only certain aspects of the child's everyday life and behaviour, and might be somewhat biased in his own methods of estimation or by his memory of his own previous ratings, an attempt was made to procure assessments from as many different observers as possible. Further, in order to determine how far the pubertal changes persisted into adult life, it was intended to supplement these earlier assessments by a similar set to be procured some eight or ten years later when each child was fully grown.

Largely as a result of the war, this latter part of the plan has not been very effectively carried out ; and, as it now appears, the results obtained in certain individual cases prove even more interesting and instructive than the generalizations from the entire group. In the present article we shall confine ourselves to one of these cases only ; and our primary object will be to explain the method adopted. In a subsequent paper it is proposed to compare the results obtained from a larger number of cases, and to review the conclusions reached by those who have been working with us along very similar lines.

A. THE CASE HISTORY

The case we shall choose will be that of a girl for whom nine successive assessments were obtained during the pubertal stage, and who, as it happened, was encountered once again quite recently as a student. In this instance, therefore, an exceptionally reliable set of adult assessments was obtainable, and in addition the girl's own retrospective comments were secured as a check on the inferences drawn.

The girl, as first seen, was a bright and well developed child, nearly 12 years of

[1] Some account of this enquiry (without the subsequent follow-up records which are still incomplete) was given in a *Report of the L.C.C. Psychologist* (' On the Mental Development of Pupils during Early Adolescence,' 1924). The enquiry itself was part of a more ambitious scheme for a longitudinal study, including physical as well as mental records, of a series of selected cases during the whole of their school career.

age. She was referred by the headmistress of the elementary school for examination and remedial treatment on account of alleged educational backwardness. Since there seemed some possibility that her backwardness might be due in part to emotional peculiarities, systematic ratings were obtained for her various temperamental traits. Several persons both at the school and at the psychological laboratory were asked to assess her at intervals of six months ; and, owing to special circumstances to be related in a moment, additional assessments were procured during the intervals. The list of ratings, and the ages at which they were made, is shown in Table I.

TABLE I. LIST OF SUCCESSIVE RATINGS

No. of Rating	Age	Ratings and Relevant Incidents
1.	$11\frac{8}{12}$	First reference and elementary school rating (class teacher)
2.	$12\frac{0}{12}$	Psychologist's first rating (C. B.)
–	$12\frac{1}{12}$	Transference to secondary school
3.	$12\frac{3}{12}$	First secondary school rating (headmistress)
–	$12\frac{4}{12}$	Onset of puberty
4.	$12\frac{5}{12}$	Psychologist's second rating (E. W. S.)
5.	$12\frac{6}{12}$	Second secondary school rating (1st form mistress)
6.	$13\frac{0}{12}$	Third secondary school rating (2nd form mistress)
–	$13\frac{1}{12}$	Attempted sexual assault
–	$13\frac{2}{12}$	Commencement of ascertained delinquencies
7.	$13\frac{3}{12}$	Psychologist's third rating (C. B.)
8.	$13\frac{6}{12}$	Psychologist's fourth rating (A. N.)
9.	$14\frac{0}{12}$	Fourth secondary school rating (3rd form mistress : H. W.)
10.	$23\frac{0}{12}$	College ratings (averaged)

As will be seen from the table, each rating, except the second and fourth, came from a different person. The observers comprised three psychologists, who were concerned with the investigation of the case at the laboratory, and five teachers, one from the elementary and four from the secondary school.

The traits assessed and compared are enumerated on the left-hand side of Table IV. All except two (emotional stability and emotional maturity) represent conative tendencies—i.e., what McDougall would call ' mental forces,' as distinct from ' mental abilities ' ; and the list is limited to what he describes as ' primary ' traits. At the earlier stages, a slightly longer list was employed [1] ; but for purposes of correlation we have omitted those which certain observers were in the end unable to assess. It will be unnecessary to set out either the definitions of the several traits or the instructions given to the assessors. Fuller information on these points can be obtained from previous publications (cf. *The Young Delinquent*, chapter on ' Temperamental Conditions,' pp. 417f. and refs.). Briefly it may be said that the assessors were asked to estimate the traits so far as possible from actual behaviour. A 5-letter scale with pluses and minuses (easily converted to a numerical scale) was provided, and its use explained. One or two methods of checking the ratings were suggested, the most important being that the observer should make a mental ranking of the traits in order of prominence. Thus, to check what may be called his absolute assessments, he was also asked to note whether the relative order appeared to him to be correct (e.g., if the child's highest score was for joy, this would naturally mean that joy was the most conspicuous feature in the child's conduct about the time of assessment ; if disgust received a higher score than sex, this would mean that the tendency to feel disgusted would be likely to prevail over the tendency to feel sexually interested, and so on). Since our interest centres rather on the correlations between the ratings than on the ratings themselves, there is no need to tabulate the detailed ratings in full : if required, these can be obtained from the original report, or reconstructed from the factor-saturations and factor-measurements. But a brief verbal account of the history of the case seems indispensable in order to understand the problems that arose.

Mary G—— was the only child of fairly well-to-do parents, living in a good residential district. The headmistress had thought her able enough to win a junior county scholarship to a secondary

[1] For example, as with all such cases, the psychologist concerned made assessments for more complex emotional traits (interests, sentiments, complexes, and the like) ; and, of course, any complete description of the conative or ' dynamic ' structure of an individual personality would inevitably have to take account of these. The way in which this may be done has been sufficiently discussed by one of us elsewhere (e.g., *The Young Delinquent*, pp. 526–68.

school. The girl, however, failed, chiefly because of her backwardness in the more mechanical subjects —an outcome which her class teacher had in fact foreseen. There was no doubt about the child's innate ability. When tested at the laboratory, she had an intelligence quotient of 134 ; and, with the exception of two supplementary tests (for attention and memory for meaningless syllables), her performance in all tests of special aptitude was well above the average for her age. As regards temperament and character, the headmistress reported that " Mary seems on the whole fairly normal, though she is rather fond of getting into the limelight, and is given to frequent and somewhat noisy giggling. Partly for this reason one teacher, who found her difficult to cope with, described her as hysterical. But I myself do not consider her lacking in emotional balance to any grave extent, though I admit that, like many healthy girls of her age, she is merry, high-spirited, and at times inattentive in the more mechanical lessons."

At the psychological laboratory, no sign of any psychoneurotic tendency could be detected. The psychiatrist, after a brief interview, reported that " both bodily and mentally she seems a perfectly healthy child ; her physical development is good, indeed, rather in advance of her actual years." While attending the laboratory she rapidly improved in emotional stability ; and eight weeks later was said to be " now bestowing far more interest and attention on her lessons." She continued to come regularly for another 12 months to receive further coaching in her weakest subjects, particularly spelling and mechanical arithmetic. When she was 12 years of age, her parents transferred her to a secondary school as a fee-paying pupil. In connexion with her entrance, the new headmistress (a shrewd judge of character who had already assisted our researches in similar ways) drew up a special set of ratings. Puberty commenced somewhat early, at the age of 12 years 4 months ; and another set of ratings was therefore almost immediately obtained.

Some months later the headmistress reported a series of reckless thefts accompanied by a marked increase in emotional instability. " The child " (she wrote) " has become strangely ill-tempered and defiant : she is extremely disobedient, and there are even rumours of [sexual] misconduct. When questioned about her misdoings, she is more inclined to parade than conceal them ; and betrays no sense of grief or penitence at all." Both the mistress and the mother attributed these unexpected changes to disturbances ensuing after the premature onset of puberty. During the next three or four months the child was interviewed almost weekly at the psychological laboratory ; and the outbreaks of misbehaviour became slighter and less frequent. At one of the later interviews she described how she had been subjected to a not very serious sexual assault by an older boy ; and shortly after this confession the thefts (or at any rate the ascertained thefts) entirely ceased, and her work in class was reported to be " showing a remarkable improvement." The last school rating was obtained when she was 14 years of age. Owing to her parents' move from London, she was seen no more while still at school. But nine years later she was encountered again as a student reading for a diploma course ; and it thus became possible to secure a fresh set of reliable ratings both from members of the teaching staff and from a postgraduate student of psychology who knew her intimately. These assessments were averaged ; and the resulting figures are inserted in the second column in Table IV.

The questions raised by this child's history seem at once typical and clear. First, how far (if at all) have (i) change of school, (ii) onset of puberty, and (iii) the incident of the attempted assault altered the child's emotional behaviour ? Secondly, when such alterations have occurred, what is their special nature ? In particular, is the change from good behaviour to delinquency associated with the effects of puberty, or with those of the later emotional shock ? And finally, how strong are the new tendencies in comparison both with each other and with the child's initial character, and to what extent have they become stable and persistent components ?

Questions like these are constantly cropping up in studies in individual psychology. But almost always the replies given rest on casual impression or personal surmise, and on very little else.

B. CORRELATIONS AND FACTOR ANALYSIS

Reliability. At the outset it will be desirable to get some general indication of the reliability of the data available for answering these questions. As it happens, much the same list of traits was used for the other children in the larger group studied for the purposes of the wider research. With rare exceptions each of the children attended a different school ; hence the groups of assessors were different in almost every case. In such circumstances, as Burt has argued elsewhere, the appropriate method of estimating the reliability of the assessments is to calculate an *intra-class*

correlation with equalized means. The coefficients so obtained for each trait are shown in the first column of Table IV. For all except four the value is over 0·75. As usual, sex has the lowest reliability of all.

Methods of Analysis. In the *Report* already cited, it was pointed out that there are two obvious ways of analysing assessments or other data which have been obtained in the way described above, namely, (i) by analysis of variance and (ii) by factor analysis. The relations between the two methods, and their relative merits for different types of problem, have already been discussed in an earlier number of this *Journal* (5 : cf. also *Brit. J. Educ. Psychol.*, XVI, pp. 129–31).

Consider first of all a single trait, emotional stability, let us say. If we desired to know whether the onset of puberty has appreciably increased or decreased this characteristic, we could average the ratings assigned for it (1*a*) before and (1*b*) after puberty had commenced, and then test the significance of the difference between the means in the usual way. Here, however, we are concerned with a whole set of traits. Hence it will be simpler to test significance by an analysis of variance, with due allowance, if necessary, for inter-correlations or covariances. In doing so, we could, if we wished, introduce more than one criterion of classification : we could cross-classify the ratings into those (2*a*) before and (2*b*) after entrance to the secondary school, or again into (3*a*) teachers' ratings based on behaviour at school and (3*b*) psychologists' ratings based on interviews at the laboratory ; and by a further analysis of variance we could then test the significance of the effects attributable to these various influences and of their interactions at different levels. This would seem to be the appropriate method where the main problem is to ascertain whether the causes envisaged by certain preconceived hypotheses are *veræ causæ* or not.

If, however, our interest lies rather in the *nature* of the effects than in their mere statistical significance, then factor analysis would seem the more suitable technique. Quite possibly the factorial analysis will lead to much the same classificatory principles. But their manifestations will now be expressed by weighted sums rather than by unweighted (i.e., ordinary means). The results thus obtained will thus give us some notion of the patterns or configurations involved. This, therefore, appears to be the more appropriate procedure to follow upon here.[1]

A factorial analysis should enable us to decide fairly easily between the main alternative hypotheses that have been put forward. Here, for example, the essential question is whether the predominant influence is the child's original constitution or the accidents of her recent life history. In the past numerous writers have maintained that emotional behaviour is mainly determined by innate endowment, and that environment serves chiefly as a provocative stimulus to elicit such behaviour, and in general produces by itself no large or lasting effects. Such a view was strongly supported by early eugenists like Karl Pearson. If it were true, we should expect to find that the only '*common factor*' in the assessments here obtained was that resulting from the stable pattern of the child's individual temperament, as manifested from the very outset : the factors arising from the different occasions on which the successive assessments were made would be, in the strict sense, '*specific factors*,' that is, they would each be peculiar to a single occasion. The resulting correlation table would then form (within the margin of random errors) a matrix of rank one, i.e., a perfect hierarchy.

If the table of correlations departs appreciably from a purely hierarchical order, then other explanations will have to be invoked. The additional factors might be regarded as factors common either (*a*) to certain groups of occasions (e.g., all those occasions following some particularly disturbing incident), or (*b*) to certain groups of

[1] There is, of course, no reason why the two methods should not be combined, as, indeed, was attempted in the earlier *Report*, where the significance of the effects as well as their nature seemed to be in question. In theory the factors themselves should be made the object of a direct test of significance. Of the statistical tests at present available for this purpose the more rigorous are decidedly laborious to apply, and are moreover strictly applicable only when certain special modes of factorization have been used (2) ; the simpler tests provide merely rough and approximate figures, but that is commonly sufficient for most purposes : after all, what we need to know is simply whether the results are significant enough to support some general conclusion.

persons (e.g., observers having the same standpoint or the same kind of contact with the child).[1]

In cases like the present the ratings obtained are ratings for the same individual (the 'subject') made by a group of different persons (the 'observers'). Now too often it is forgotten that the data gathered as a result of a rating process consists of responses of one or more minds (those of the observers) to a highly complex situation in which the particular mind observed (that of the person to be rated) forms only one element. Hence the ratings submitted really constitute a portion of the behaviour of the observers, even when we seek their co-operation with a view to gaining light on the behaviour of someone else. Occasionally, indeed, in certain enquiries, the purpose of 'correlating persons' is actually to investigate the differing attitudes of the 'observers'; and in such cases the 'subject' of the ratings may himself be of no interest whatever, except as furnishing a kind of common stimulus or a sample of concrete test-material to elicit reactions from the observers. In the present instance, however, the aim of the investigation (as is more usually the case) centres on a study of the subject. Accordingly, as mentioned above, special devices have been used in the hope of eliminating personal differences between the observers, and so making the rating process as objective as possible. Further, by appropriately selecting the observers, an attempt was made both to neutralize individual bias and at the same time to provide data for estimating how far (if at all) the personal attitudes or prejudices of each one had in fact still affected his judgements. We shall therefore have to examine with what success these irrelevant influences may have been excluded; and this again can best be done by comparing the correlations.

Correlations. The correlations [2] between the nine ratings obtained during childhood are set out in Table II.

The pattern presented by the correlation table is itself instructive. To begin with, it is by no means a perfect hierarchy. For the most part the highest coefficients in each column or row seem concentrated in the cells adjacent to the diagonal; and there is a tendency for moderately high coefficients to reappear at or near the top right-hand and bottom left-hand corners. In other words, the table exhibits clear

[1] The factors relating respectively to different Persons, different Occasions, and different Traits are for brevity sometimes referred to as '*P*-factors' '*O*-factors,' and '*T*-factors.' Similarly the process of factorizing assessments or trait-measurements, when the correlated variables are *persons* (e.g., different examiners, observers, or judges) has sometimes been called '*P*-technique,' while the process of factorizing assessments obtained for the same person on successive *occasions* has been called '*O*-technique.' As has been explained elsewhere, these and similar shorthand phrases ('*R*-technique,' '*Q*-technique,' '*S*-technique,' '*T*-technique,' and the like) have been used colloquially when talking over projects with research-students in the laboratory. But, as one of us has already pointed out, the idea that statistical psychology consists in the mechanical application of specific 'techniques,' attractive as it is to the elementary student, seems particularly misleading in the study of individual personality. Moreover, the designations themselves are not very informative in published works intended to be read by workers in other laboratories where the same 'family slang' may not obtain, and may even engender some confusion. Thus, Cattell and Stephenson (both former students of the laboratory) have each since defined these informal labels in his own particular way. In one passage, for example, Cattell (9, p. 98f.), quoting Burt's *Factors of the Mind*, uses the phrase '*P*-technique' to denote 'correlation between persons': but in a footnote elsewhere he says "it would be more convenient to limit Burt's term *P*-technique to analyses with a single person" (Burt's 'intra-individual' correlation: cf. 3, p. 60 *et seq.*). This latter procedure has more frequently been termed '*O*-technique,' at any rate when (as is nearly always the case) the single person is assessed on different *occasions*. But, so far as method is concerned, when (as most frequently occurs) the different assessments are themselves obtained from different persons,'*P*-technique' and '*O*-technique' are really the same; and whether we call the factors *P*-factors or *O*-factors is a question, not of statistical 'technique,' but subsequent interpretation.

When we are comparing a number of assessments on successive occasions, considered in chronological order, the problem of *trend* is likely to arise. In such cases, the modified forms of factor analysis that have recently been developed by statistical economists will be found of special value: (cf. 11 and 13). In the present case no definite trends are discernible. But with other children we have found such procedures perfectly practicable, though somewhat laborious. And one of us (H. Watson) hopes shortly to publish an illustrative research where this further procedure has been adopted.

[2] It should be explained that Miss Watson completed practically all the calculations here described soon after the last series of school ratings was secured; and most of the inferences drawn in the text were already formulated in a brief report on the case, long before any of the later confirmatory evidence was secured.

TABLE II. CORRELATIONS BETWEEN RATINGS

Rating	Age	1	2	3	4	5	6	7	8	9
1	$11\frac{8}{12}$	(·661)	·612	·535	·391	·415	·520	·561	·495	·624
2	$12\frac{0}{12}$	·612	(·522)	·463	·507	·469	·427	·384	·531	·547
3	$12\frac{3}{12}$	·535	·463	(·456)	·486	·472	·518	·358	·523	·536
4	$12\frac{5}{12}$	·391	·507	·486	(·761)	·681	·754	·697	·738	·635
5	$12\frac{6}{12}$	·415	·469	·472	·681	(·666)	·663	·566	·621	·594
6	$13\frac{0}{12}$	·520	·427	·518	·754	·663	(·682)	·672	·713	·641
7	$13\frac{3}{12}$	·561	·384	·358	·697	·566	·672	(·985)	·949	·670
8	$13\frac{6}{12}$	·495	·531	·523	·738	·621	·713	·949	(·934)	·716
9	$14\frac{0}{12}$	·624	·547	·536	·635	·594	·641	·670	·716	(·674)

evidence of what one of us has elsewhere termed 'cyclic overlap.'[1] (There are a few suggestive exceptions, nearly all of them among the figures for the 4th rating.)

As with all 'correlations between persons,' the coefficients so obtained might themselves be described as 'reliability coefficients,' at least in one sense of that equivocal term; but such a description is, in our view, apt to be misleading.[2] Judged by the customary standard for such coefficients, some of the figures are decidedly low; three are less than 0·400. These low values appear, not only when the ratings are separated by fairly long intervals of time (the lowest of all is for ratings separated by a good 12 months), but also after one or other of the more conspicuous changes to which we have referred. In fact it is the very occurrence of these low correlations that supplies the objective evidence for some new change in the child's emotional attitude. To escape the inevitable reduction that occurs with all such coefficients as a result of mere passage of time, we may restrict our comparison to adjacent ratings. For these the average interval is little over three months; and the correlations, with two exceptions, are all above 0·600, and average 0·765. The two exceptions relate to assessments obtained, first after removal to the secondary school, and secondly after the onset of puberty.

As we have seen, the 2nd, 4th, 7th, and 8th ratings were obtained from psychologists—the 2nd and 7th, indeed, being from the same individual, and the 3rd, 5th, 6th, and 9th from mistresses at the secondary school. The inter-correlations between observers from the same group (averaging 0·634 for the psychologists and 0·571 for the teachers) do not differ appreciably from the cross-correlations between members of different groups (average 0·582); and the correlation for the two sets of ratings given by the same person (0·384) happens to be one of the lowest in the table. So far, then, there appears to be little or no sign of subjective bias—either professional or

[1] Burt, C., *L.C.C. Report on Distribution and Relations of Educational Abilities*, 1917, p. 60. As is there noted, the phenomenon is especially apt to occur in correlations between marks or measurements procured for the same person or persons with the same set of traits or tests *on successive occasions*.

[2] This usage lends a fallacious support to the common notion that the estimation of reliability is irrelevant to, and indeed impossible in, the assessment of orectic or 'dynamic' traits. Miss Ainsworth, for example, states that "in the context of psychometric (i.e., cognitive) tests, a test cannot be valid without being reliable"; this, she says, "implies that the function being measured is a stable function"; but, if "the dynamic structure of an individual's personality" is subject to change (as everyone would, of course, expect), then, she maintains, it cannot be stable, and hence cannot be assessed for reliability (1). However, the alleged implication is erroneous. The 'reliability' of a set of measurements (as is now almost universally agreed) is defined by the ratio of the true variance to the error variance (cf., for example, Guilford, J. P., *Fundamental Statistics in Psychology and Education*, 1950, p. 474). If a variable changes with time, that no doubt will prevent repeated testing at wide intervals being used to estimate the reliability of the test employed; but it does not, for instance, prevent the use of simultaneous (or nearly simultaneous) assessments. Otherwise we could never measure the reliability of a watch. There is, in short, a world of difference between a stable test for an unstable trait and an unstable test for a stable trait.

It is true that Ferguson, referring to the work of Burt and others on correlating persons, has suggested that we might speak of the 'reliability of the persons' in contrast to the 'reliability of the tests' (or other method of assessment), where, by 'reliability of persons,' he means the reliability of the single persons *observed*, not the reliability of the persons who *make the observations* (*Reliability of Mental Tests*, p. 9). But the use of the same word to denote very different things is surely to be deprecated.

individual—on the part of the observers. However, it is not altogether fair to trust to inspection of the raw coefficients, whose size must depend mainly on more powerful influences. Let us therefore examine what is left after the first three factors have been eliminated. We then discover a few small positive residuals, whose nature is certainly consistent with the existence of some slight personal bias. Yet the amounts observable remain far too slender to be statistically significant. Hence, in this enquiry at any rate, we may safely conclude that the personality of the different observers has had little or no influence, assuredly none large enough to distort the answers to our main problems.

A. *General and Bipolar Factors.* A factor analysis by simple summation was carried out first of all. After three factors had been extracted, every residual was found to be numerically less than 0·06. The factor-saturations are shown in Table III A.

The first factor evidently represents the average of the child's emotional behaviour during the entire period under review ; it accounts for over 58 per cent. of the total variance. The second factor accounts for 7 per cent. ; it divides the period roughly into two main sections : first, the portion before the onset of puberty and, secondly, the portion after it, or rather the portion during which the child's behaviour was disturbed, since the last rating of all suggests a reversion to the more stable type of conduct which dominated the picture before any disturbances commenced. The third factor accounts for 4 per cent. of the variance : it divides the disturbed portion into two subsections, first, that which apparently shows the immediate effects of puberty and, secondly, that which seems to indicate the effects of the assault. The remaining saturations for this factor are too low to claim any significance : but it is noteworthy that the pre-pubertal ratings (2 and 3), though numerically small, have the same sign as the pubertal.

B. *Triangular Group Factors.* Instructive as they are from a purely statistical standpoint, these summational factors do not afford a very suitable way of representing the causal processes that are presumably at work. It would be more natural to treat the effects of puberty and of the sexual incident as something added to, or superposed upon, the child's basic personality. Hence an analysis in terms of positive group factors should furnish a far less artificial picture of the component tendencies than the preliminary summational analysis into general and bipolar factors.

Both the pattern displayed by the table of correlations and the nature of the bipolar sign-pattern suggest that a triangular type of group factor matrix is likely to yield the closest fit. The distribution of plus and minus signs and non-significant figures in Table III A indicates the probable limits to the several factors. Accordingly a preliminary factorization was carried out by the triangular method : (for the working procedure adopted, see 8, pp. 40–4). The summational factors were then arithmetically rotated to fit this general group factor scheme (for the method of rotation, see 7, p. 75). The results finally reached are set out in Table III B.

TABLE III. FACTOR-SATURATIONS

Rating	Age	A. Summational Factors I′	II′	III′	B. Group Factors I	II	III
1	$11\frac{8}{12}$	·695	—·437	—·091	·826	—·053	·042
2	$12\frac{0}{12}$	·646	—·313	·027	·724	·061	—·075
3	$12\frac{3}{12}$	·629	—·208	·116	·652	·170	·033
4	$12\frac{5}{12}$	·818	·241	·236	·578	·629	·122
5	$12\frac{6}{12}$	·737	·121	·293	·593	·556	—·084
6	$13\frac{0}{12}$	·809	·136	·095	·635	·533	·069
7	$13\frac{3}{12}$	·835	·334	—·437	·564	·525	·634
8	$13\frac{6}{12}$	·902	·220	—·188	·621	·478	·546
9	$14\frac{0}{12}$	·815	—·094	—·031	·745	·297	·162
Square Sum..		5·338	·592	·401	3·964	1·627	·762
Contribution to Variance ..		59·3	6·6	4·4	44·0	18·1	8·5

The inferences to be drawn from the triangular factor-matrix thus obtained are fairly obvious. The first factor has positive saturations for all the nine ratings. It is therefore what I have elsewhere termed a ' basic ' factor. It accounts for about 44 per cent. of the total variance. Its influence persists throughout the whole time that the child was under observation ; and, though diminishing for a while to a slight extent, it emerges almost as dominant as ever in the last rating of all.

The second factor is a group factor with non-significant saturations in the first two sets of ratings. It accounts for 18 per cent. of the total variance. Since its saturations are highest after the onset of puberty, it may be presumed to represent the consequences of the pubertal change. We may note, however, that, judging by the unexpected figure of ·170 in the second column in Table III B, there is some possibility that the new tendencies may actually have appeared shortly after the child's removal to the secondary school and before the appearance of any physiological signs of puberty. This could be explained by the fact, noted in other connexions, that not infrequently the mental changes associated with adolescence may precede the physical manifestations. On the other hand, it might really be due, not to internal changes at all, but rather to the child's new association with older and more mature girls. These alternatives, however, present a subsidiary problem which cannot be answered by examination of the present data.[1]

The third factor is a small group factor with its highest saturations confined to the last three ratings, i.e., those reported after the sexual incident already referred to. It accounts for barely 9 per cent. of the total variance. But it is perhaps worth noting that similar manifestations seem to make a slight temporary appearance as early as the 4th rating, i.e., immediately after the onset of puberty ; while in the last rating of all its influence already shows a large decline.

The figures given for the contributions to the variance relate to the whole period under review. If, however, we compare the figures for the latter portion of the child's thirteenth year only (ages $12\frac{5}{12}$ to $13\frac{0}{12}$ inclusive), we shall find that factor II (the pubertal factor) contributes for those few months almost, but not quite, as much as factor I (the basic factor). And if we then compare the figures for the six months following the sexual assault, we shall see that factor III (representing the effects of the assault) contributes as much as factor I, while factor II contributes far less—only about two-thirds of either of the other two.

C. THE FACTOR-MEASUREMENTS

The Psychological Nature of the Factors. To determine the precise nature of the three factors thus found we must turn from the factor-saturations to the factor-measurements. These have been calculated in the usual way, by first computing the regressions and then taking weighted sums of the original ratings. The results, in unitary standard measure, are shown in Table IV. We venture to think that these figures present a very clear reflection both of the child's basic personality and of the changes that ensued after the two incidents we have referred to, namely, the onset of puberty and the sexual assault.

Factor I. The first factor differs very little from the initial ratings sent in by the first two observers. It depicts what we have described above as the child's basic personality. To use the language of the

[1] We may add that (quite apart from other evidence) the factorial analysis of later cases leaves no doubt that both influences may operate. Certainly, many of the mental changes commonly regarded as the intrinsic and spontaneous manifestations of physiological pubescence may be produced by the alteration in external social relations which so often take place at about the same time : thus, in children who leave school for work at an early age, they appear early ; among students who remain *in statu pupillari* till 20 or 21, they may be delayed until much later.

temperamental types to which more general factorial studies have led (4, 6), she is a highly 'euphoric' and a moderately 'extravert' personality. At this stage the most outstanding feature on the positive side was her joyful gaiety. Her next most prominent tendencies were sociability and talkativeness. But her self-assertive impulses, which, in the very first set of teacher's ratings, were assessed as much stronger than self-submission, seemed rapidly to decline. In the main she appears to have been considered an obedient, reasonably submissive, and even affectionate child. Sex received a low positive rating : in this respect she was evidently regarded as a fairly average girl for her age. On the negative side the most striking features were her unusual freedom from fear and nervousness, her unfailing good temper, and her marked immaturity : her emotional demeanour, indeed, was rather that of a child 18 months younger. The first set of ratings graded her rather low for stability : but in this, too, she rapidly improved. Partly no doubt because of her parents' comfortable circumstances, she had little reason to covet other people's goods, and she was at this stage habitually generous with her pocket-money, her playthings, and her other possessions : hence the low rating for acquisitiveness. Disgust and sorrow also seem to be weak and unimportant tendencies ; for these the figures, though negative, are small, and here, too, therefore, she is probably to be considered an average child.

TABLE IV. FACTOR-MEASUREMENTS, RELIABILITY COEFFICIENTS, AND FINAL RATING

Trait	Reliability	Factor I	Trait	Factor II	Trait	Factor III	Age 20
Joy	·92	·412	Sex	·424	Disgust	·378	·36
Sociability	·87	·346	Affection	·353	Anger	·325	·39
Talkativeness	·94	·300	Sociability	·268	Acquisitiveness	·290	·23
Self-submission	·81	·197	Maturity	·234	Self-assertion	·228	·18
Affection	·77	·178	Talkativeness	·218	Maturity	·185	·16
Self-assertion	·86	·136	Fear	·123	Talkativeness	·123	·31
Sex	·63	·112	Anger	·067	Sociability	·105	—·21
Sorrow	·76	—·112	Disgust	—·047	Joy	·079	—·18
Stability	·82	—·131	Sorrow	—·134	Affection	—·132	·30
Disgust	·72	—·136	Self-assertion	—·179	Sex	—·176	—·38
Acquisitiveness	·69	—·150	Stability	—·268	Self-submission	—·308	—·26
Maturity	·74	—·337	Self-submission	—·312	Fear	—·334	·12
Anger	·85	—·365	Joy	—·343	Stability	—·342	—·42
Fear	·79	—·450	Acquisitiveness	—·404	Sorrow	—·421	—·31

Factor II. The second factor may be regarded as representing a pubertal or adolescent personality, superimposed on the first. This mode of expression must not be taken to imply that there are two distinct and concrete persons in one and the same child. So far as her actual behaviour is concerned, the second factor might perhaps better be considered as a complex influence tending to increase or to diminish, in differing degrees, the strength of certain tendencies already pre-existing. On the positive side the most conspicuous feature is the high figure allotted to sex. That for affection is almost as large. There are positive increments for sociability and talkativeness, though here the added contributions supplied by the new factor are smaller than the basic contributions of the old. One striking change is that maturity now receives a positive instead of a negative rating. So do fear and anger, though the amounts are comparatively small. On the negative side perhaps the most remarkable difference is shown by the factor-measurement for joy : with the first factor, as we have seen, joy obtained the largest positive figure of all ; it now receives one of the largest negative values ; the effect is to annul almost entirely the influence of the previous basic factor. Self-submission, which formerly received a small positive figure, now receives a large negative : as a result, from being docile and demurely submissive, the girl has swung over to the opposite type : there is indeed no return to the old laughter and giggling, but she has become both disobedient and self-willed. Most significant of all, perhaps, is the marked negative addition to the negative figure previously allotted for emotional stability.

Factor III. Judging by the factor-saturations allotted to different dates or ages, the third factor, as we have seen, must represent the after-effects of the shock occasioned by the sexual assault ; and its characteristic contributions seem fully consistent with that interpretation. On the positive side, its most conspicuous features are the high positive measurements assigned to disgust and anger. These are tendencies which such an incident might easily waken, and they are slow to die down. Next after these is the large figure for acquisitiveness : here no doubt the assessors have been influenced by their knowledge of the child's recent thefts. Self-assertion receives a further positive rating and

so does maturity. Sex, on the other hand, now obtains a negative figure ; and the further reduction for fear and self-submission is worth noting. On the negative side, however, the largest figure of all is that assigned to sorrow : this means, not that the child is at length restored to her original gaiety (for joy gets a figure of approximately zero), but that, in the view of those who assessed her, she showed during this latest phase an undue immunity from grief, penitence, and tears. The negative increment for stability is almost as striking. If we add the contributions for stability made by all three factors, the total comes to $-\cdot741$ (the only other total approaching it is that for sorrow $-\cdot661$). Thus at the time of these last three ratings the child was in a highly depressed and unstable state.

It will be noted that the changes introduced by this last factor appear on the whole to have overtaken the child like a passing wave. As Table III B indicates, by the date of the final school rating, the saturation coefficient for factor III had dropped from $\cdot546$ to $\cdot162$. How much the decline of this third component in the child's behaviour is due to psychological treatment, and how much to the simple passage of time, we need not here discuss. The interesting feature of the case is that, as the last row of saturations shows, by the age of 14 the child had already largely reverted to what we may look upon as her normal personality. The effects of puberty and the sexual assault have perhaps not wholly disappeared ; but they have plainly become much smaller than before.

Judging by the child's actual conduct and her own later interpretations, her reactions at each stage were determined by the *balance* of tendencies, not by their absolute strength taken in isolation. The attempt at prediction is rather like the determination of a resultant movement in physics, when the component forces are known. Thus the factor, as expressing a pattern, yields a better interpretation of past behaviour, and a better basis for forecasting future behaviour, than the raw ratings, however accurate.[1]

We have so far been concerned almost exclusively with one short phase of the child's entire life, namely, with what may be called the years of pubertal change. As it happens, however, we also possess a certain amount of comparable data relating to her earlier and her later life.

The last column of Table IV gives the average ratings which (as mentioned above) were obtained for the girl at the age of 23. It will be seen that they show a far closer resemblance to the hypothetical measurements for factor I (the ' basic personality,' as we have provisionally termed it) than to those for factors II or III, which represent later stages in the child's development : the correlations in fact are (I) $+ 0{\cdot}74$, (II) $- 0{\cdot}09$, and (III) $- 0{\cdot}20$ respectively. These values confirm the tentative inference already drawn, namely, that with this girl the disturbing effects of puberty and of the subsequent emotional shock were transitory.

It may also be of some interest to consider how the same three factors are related to the child's personality during her earlier life. It is a common hypothesis, particularly among medical writers, that both puberty and emotional traumata are apt to produce a relapse to a previous and more primitive type of behaviour. In the hope of obtaining some light on this view, we obtained from the mother two sets of ratings for the child, based on her memory of the child's behaviour at the ages of about (*a*) 4 years and (*b*) 10 years. The correlations of these two ratings with the three factors are (*a*) $+ 0{\cdot}16$, $- 0{\cdot}26$, and $- 0{\cdot}19$, and (*b*) $+ 0{\cdot}31$, $- 0{\cdot}41$, and $- 0{\cdot}28$ respectively. So far as they go, these figures plainly tend to contradict rather than to support the notion that the incidents mentioned were followed by an infantile regression.

Conclusion. Even if the data analysed are not always as reliable as might be wished, it will now, we think, be abundantly clear that the method of statistical analysis we have employed will enable investigators to answer, on the basis of objective evidence, questions like those propounded at the outset of this discussion. In the

[1] This was notably the case with the last factor, where the essential figures were deduced before a direct analysis of the child's apparent motives was reached during the final interviews. It is noteworthy that the figures are indicative of tendencies which may be more or less unconscious or repressed, e.g., the high saturation for sex in the second factor, which was based on data secured before the alleged sexual ' assault.' In the main the interpretation given in the paragraphs above was fully confirmed by discussions some ten years later, when the girl could look back on her own behaviour more objectively. According to her own account, however, she had been by no means so completely ' unconscious ' of the underlying impulses as her reticence had led us to assume at the time.

present instance, it will be remembered, the immediate problem was to explain the child's unexpected spell of stealing and ill-temper : two conceivable hypotheses were suggested : first, that it was due to the premature onset of puberty ; secondly, that it was the consequence of an emotional shock—a revulsion from the sexual incident already described. The figures obtained leave no doubt that the second explanation is in the main correct. Further, considerable light has incidentally been thrown on more general problems. The analysis shows that, in the case of at least one child, the change of school and the advent of puberty, as well as the incident just mentioned, have exercised a definite influence on her subsequent emotional behaviour, although in very different degrees. The proportions calculated for the contributions to the total variance serve to indicate the relative importance of these causal agencies, both as compared with each other and as compared with the child's previous behaviour ; and the pattern formed by the factor-saturations depicts the concrete nature of the changes attributable to the several causes.

The supplementary data, comparing the factors found with earlier and later ratings, are admittedly unreliable, particularly those based merely on the mother's recollections. Moreover, the conclusions here reached relate solely to one particular case. Nevertheless, with the assistance of other investigators, we have now collected a considerable amount of information in respect of other children, many of them studied at first hand from a much earlier age. Among those who happened to have been subjected to a similar kind of stress or strain, the same method of analysis has shown that, in an appreciable proportion at least (about 27 per cent.), something very like a temporary regression to the behaviour of an earlier age did in fact ensue. Further, within this smaller batch two distinguishable types or tendencies may be discerned : first, those whose emotional behaviour is especially liable to be changed by internal developments (such as puberty), and secondly, those whose behaviour is especially liable to be changed by external incidents (such as an emotional shock or crisis). One of us hopes to publish fuller evidence on these points in a later paper, with analyses and after-histories, not of a single case, but of a series. In this introductory note our main object has been merely to show how a factorial procedure may help to supply objective evidence on many of the familiar problems of individual personality, and to illustrate in some detail the working methods which we and our fellow-workers have employed.

III. SUMMARY

1. In individual psychology a neglected mode of investigation consists in applying statistical analysis to the study not of groups but of single persons, i.e., to the study of what in an earlier paper (3, p. 60) was called ' intra-individual ' as contrasted with ' inter-individual ' variations.[1] In this way, conclusions reached by the impressionistic methods of so-called ' clinical observation ' can often be subjected to more rigorous empirical tests. The chief methods available are analysis of variance and factor analysis. With the latter the technique to be used is a special case of that adopted in correlating persons as distinct from traits.

2. A concrete example is described in detail to illustrate how the components due to internal developments (such as maturation or the onset of puberty) and to external influences (such as change of social environment or emotional shocks and strains) can be isolated, and expressed as a numerical configuration of traits. The method used consists in calculating a rotated matrix of group factors (here a triangular matrix) with the aid of a preliminary analysis by simple summation.

[1] The phrase was taken from Stern who distinguishes two main types of ' intra-correlations '—*Übungs-Korrelationen* and *Entwickelungs-Korrelationen* (*Differentielle Psychologie*, pp. 152, 289–90). In this paper we have been concerned with the second type.

3. The results obtained show that the factor-saturations so computed give a clear picture of the superadded changes due to particular incidents or developments in the individual's life ; the factor-variances express the relative influence of each ; and the factor-measurements indicate the specific pattern of traits or tendencies that has been thus produced. Thus, provided the conditions necessary for a valid factorial analysis are approximately satisfied, the procedure is able to furnish objective evidence on many disputed problems in the field of individual psychology.

4. The results of the factorial analysis were briefly described in an earlier report. But the case selected seemed to deserve a fuller discussion owing to the fact that further data, procured after an interval of nearly ten years, has furnished unexpected confirmation of the inferences then drawn. This confirmation should in some measure answer the criticisms originally urged against our application of factorial methods to correlations of this type.

REFERENCES

1. Ainsworth, M. D. (1951). ' Some problems of validation of projective techniques.' *Brit. J. Med. Psychol.*, XXIV, 151–61.
2. Bartlett, M. S. (1950). ' Tests of significance in factor analysis.' *Brit. J. Psych., Stat. Sect.*, III, 77–85.
3. Burt, C. (1937). ' Correlations between persons.' *Brit. J. Psych., Gen. Sect.*, XXVIII, 59–96.
4. Burt, C. (1938). ' The analysis of temperament.' *Brit. J. Med. Psychol.*, XVII, 158–88.
5. Burt, C. (1947). ' A comparison of factor analysis and analysis of variance.' *Brit. J. Psych., Gen. Sect.*, I, 3–26.
6. Burt, C. (1948). ' The factorial study of temperamental traits.' *Brit. J. Psych., Stat. Sect.*, I, 178–203.
7. Burt, C. (1950). ' Group factor analysis.' *Brit. J. Psych., Stat. Sect.*, III, 40–75.
8. Burt, C. (1951). ' The numerical solution of linear equations.' *Brit. J. Psych., Stat. Sect.*, IV, 31–54.
9. Cattell, R. B. (1946). *The Description and Measurement of Personality.* New York : World Book Company.
10. Davies, M. (1939). ' The general factor in correlations between persons.' *Brit. J. Psych., Gen. Sect.*, XXIX, 404–21.
11. Geary, R. C. (1949). ' Determination of linear relations between systematic parts of variables with errors of observation the variances of which are unknown.' *Econometrica*, XVII, 30–58.
12. Spearman, C. (1934). ' The battle between intuitionists and psychometrists.' *Brit. J. Psych., Gen. Sect.*, XXIV, 403–7.
13. Tintner, G. (1946). ' Multiple regression for systems of equations.' *Econometrica*, XIV, 5–21.

BOOK REVIEWS

The Mathematical Theory of Communication. By C. E. Shannon and W. Weaver. Urbana : University of Illinois Press. 1949. 19*s*.

This volume comprises two contributions. The second, by Warren Weaver, consists of a non-technical introduction to the general subject of ' analytical communication ' (as the writers term it), and might well have been placed first. The word ' communication ' is used in a very broad sense. It " includes all the procedures by which one mind may affect another," i.e., not only oral speech and written messages, but " all human behaviour " : and is even extended to cover the various devices whereby the behaviour of an inanimate object (say, a moving aeroplane) is made to communicate information to a receptive mechanism (part of a guided projectile perhaps, sent to chase the first object) in such a way that the reactions of this second object are continuously adjusted to comply with the movements of the first, and so eventually achieve some prescribed purpose.

The study of communication as thus defined involves, so we are told, problems at three different levels : (i) the technical problem—how accurately can the symbols of communication be transmitted ? (ii) the semantic problem how accurately do the transmitted symbols convey the intended meaning ? (iii) the effectiveness of the problem—how effectively does the received meaning influence conduct in the way desired ? At first sight it might be supposed that the mathematical theory of the engineering aspects, as propounded by Shannon and his colleagues, applies to the first type only. But Weaver endeavours to show that, to a large extent, the investigation of the second and third types of problem consists of corollaries drawn from the theory of the first. If so, the theory would be of the greatest importance for the psychological study of language and other forms of human communication.

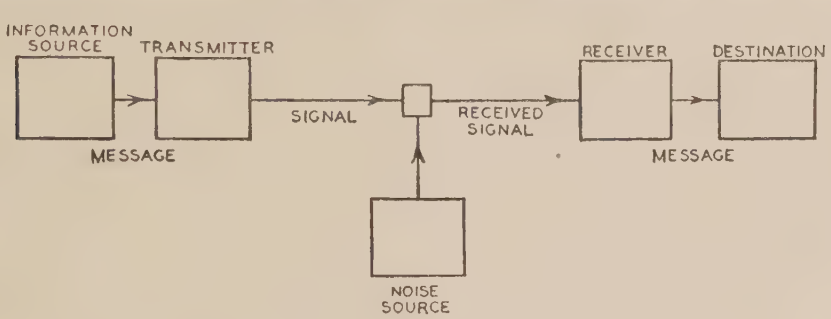


Fig. 1. Schematic Diagram of a General Communication System

From a technical standpoint the fundamental problem is this : given a message which has been selected at a first point, how can we best reproduce it at a second point ? The chief phases in the processes thus involved may best be understood from the above diagram, which is given by both Weaver and Shannon (see Fig. 1). The scheme may be regarded as a generalization of the familiar diagrams given in text-books of elementary physics to show how spatial patterns may be projected by a suitable optical device and rendered visible on a screen (as in the ordinary projection lantern), or how audible patterns may be conveyed, with or without wires, to a receiver (as in ordinary telephoning and broadcasting). All these are mechanisms for communicating information ; and, if efficient, they should function in such a way that a minimum amount of the information embodied in the patterns is lost. Moreover, as Weaver explains, the same formal principles should remain applicable when we interpret communication in a far more general sense. Burt, for instance, in a recent contribution,[1] has shown how well the Shannon-Weaver scheme will represent the processes involved in assessing psychological characteristics by means of mental tests or other devices, and

[1] This *Journal*, IV, pp. 31–7, and other papers.

has pointed out how superior such a representation is to the familiar analogy of ' projection,' when we come to enquire how the efficiency of our so-called ' projection tests ' can be studied and improved. However, as each of these writers seems to admit, psychologist as well as engineer, before any such methods can be exploited to the full, they will first have to be extended to include simultaneous patterns (like visible pictures) as well as temporal patterns (like speech and other audible codes).

In all such cases what is communicated is technically termed ' information.' ' Information ' therefore becomes the central concept in the Shannon-Weaver theory. " Unlike the concept of ' meaning,' this applies," we are told, " not to the individual message, but to the situation as a whole " : since, in any particular process of communication, the actual message to be conveyed is one chosen from a set of possible messages. Hence the selection of the message, which in turn implies the selection of the symbols that constitute that message, can be regarded as an event to which a certain probability can be attached. Thus, if we know beforehand that the voice of Russia always answers ' No,' then, when on any particular occasion that voice gives a negative reply, the amount of information conveyed is exactly *nil*. As the probability of a message diminishes from unity downwards, so the amount of information rises from zero upwards ; moreover, while probabilities combine by multiplication, the corresponding amounts of information should combine by addition. And, as Weaver points out, all these requirements will be fulfilled if we measure the information conveyed by a given event by stating the negative logarithm of the corresponding probability.

Shannon's paper carries the argument somewhat further, and elaborates it in full technical detail. His first concern is to derive the most appropriate formulæ for representing the amount of information conveyed under various conditions. He begins with the case of discrete noiseless systems and develops the well-known formula $H = -\Sigma p_i \log p_i$. The further deductions are simple enough so long as we assume that in any given message the occurrence of the successive symbols is quite independent of what precedes or follows.

But, as every enthusiast for crossword puzzles is well aware, with a given number of symbols (say, the 26 letters of the English alphabet) we never have an absolutely free choice in picking out the sequence of symbols which constitute an intelligible message. Thus, if the first letter in the message is *q*, the second will almost certainly be *u*, and the third is three times as likely to be *a* as *o*. In fact, in speaking or writing, only about half the sounds or letters we select are open to our free choice ; the rest (though we seldom recognize it) are predetermined by the statistical structure of the language. And, of course, the same would hold good of the responses or reactions that our subjects make to our test-situations : the probabilities for the different alternatives are not only unequal in themselves, but usually dependent in various ways on the responses already given.

Now, as Shannon observes, " stochastic processes of this type are known to mathematicians as ' discrete Markoff processes,' and have been extensively studied in the literature." We are thus led to elaborate the preceding formula, and put $H = -\sum_j \sum_i P_j p_j(i) \log p_j(i)$, where P_j is the probability that the jth combinational state will occur, and $p_j(i)$ is the transitional probability that, when the jth state has occurred, the ith symbol will follow. Among the various forms of Markoff process, the ' ergodic ' subtype becomes of special importance in the theory of communication.[1] And accordingly Shannon goes on to construct a number of suggestive graphs illustrating the various ergodic processes to which ' sources ' of different kinds may give rise. Then, still keeping to the ' noiseless ' type of channel, he goes on to formulate a series of theorems describing the operations of coding, decoding, and the like.

In actual practice, of course, the very process of communication involves interference with the original message, and a consequent loss of information. In communicating audible messages by telephonic methods, this interference takes the form of obscuring noise ; and it proves convenient to use the same term to cover any kind of blurring due to random interference. Shannon's second chapter, therefore, is concerned with communication by channels or media that involve the production of ' noise ' in this extended sense. In particular he examines the conditions which render possible a correction for such ' noise ' by ' filters ' and similar devices, and attempts to express in precise mathematical terms the theorems that govern the extent to which a correction can be successfully achieved.

The third and longest section of his paper is concerned with ' continuous information,' i.e., " the case where the signals or the messages or both are continuously variable." To a large extent the formulæ now required can be derived from the corresponding discrete case by a limiting process. But there are, as Shannon shows, a number of new effects which emerge in the study of ' continuous information ' and require special consideration.

[1] The term ' ergodic ' (which means ' work-path ') was introduced into statistical mechanics by Boltzmann and freely used by Willard Gibbs. It is used to designate the fact that, by a sufficient repetition of the same operation, probabilities, which were at the outset quite arbitrary, tend to become regularized. The relevant theorems, however, were only established some time later by Birkhoff, Koopman, and von Neumann. For a clear account of the matter from the standpoint of communication theory, see Wiener's *Cybernetics*, chap. II.

So far as the psychologist is concerned, the obvious and immediate interest of these various mathematical discussions lies in their application to the study of speech and speech-sounds. They enable us to measure the actual loss of information in linguistic communication whether that loss be due to the nature of the language used or to the equipment employed to facilitate communication. However, as has been suggested in the paper already quoted, this type of approach would seem to possess far wider implications. It is an essential aim, both of the statistical psychologist and of the experimentalist, to preserve as much as possible of the information contained in the data or in the sources of the data. Perhaps the clearest cases in which such principles can be effectively applied to secure this aim arise in the field of personality testing. Here, as has been pointed out, " the ' latent content ' conveyed by the testee's responses to such a test as the Rorschach constitutes the ' information ' which the testee, as a source, is communicating to the tester ; the ' superficial content ' is regarded as ' information ' plus ' noise,' and the psychologist's aim is, by improving either his experimental technique or his mathematical technique or both, to minimize the effect of the interferences and to maximize the amount of procurable information : in a word (to adopt Shannon's phraseology) to ' filter out the noise.' " Other examples could be found among the many unsolved problems arising in the study of human groups. Indeed, all social behaviour is, in one of its aspects, a process of communication, whether unconscious or deliberate ; and hence in social as in individual psychology there would seem to be a wide scope for the tentative employment of these new techniques.

W. D. S.

The Human Use of Human Beings. By Norbert Wiener. London : Eyre and Spottiswoode, 1950. Pp. viii + 242. 18*s.* net.

Three years ago Professor Wiener published a small volume entitled *Cybernetics : or Control and Communication in the Animal and the Machine.*[1] In it he developed, with the aid of somewhat abstruse mathematical arguments, what he claimed was a ' new point of view ' [2] about the processes of communicating information. He has since received, so he tells us, numerous appeals for a less technical account ; and his latest book is the answer to these requests. In writing it, his aim has been to " avoid, so far as possible, algebraic symbolism, and to emphasize the not inconsiderable social consequences of these theories."

His opening chapter, as its heading indicates, deals with the preliminary question ' What is Cybernetics ? ' The term is derived from the Greek word (used by Plato) for the art or technique of steering or directing : from the same root, through the Latin *gubernator*, we obtain our own word ' governor.' We are all familiar with the working of the ' governor ' in Watt's steam engine. The function of such an auxiliary mechanism is to ' communicate,' as it were, to the machine, by some kind of feed-back, the requisite information about its momentary state, so that, when necessary, this state can be at once corrected or controlled, and the essential purpose of the machine's activities more efficiently fulfilled. As Wiener at once points out, our own sensori-motor processes are full of examples of such ' governance,' the afferent organs forming auxiliary mechanisms for communicating information to the efferent. He goes on to explain how, as Professor of Mathematics at the Massachusetts Institute of Technology, he has been occupied with the theory of ' communication engineering,' and how, during the past twenty years or so, work in this field has led to " a new and fundamental revolution in technique," and to the construction of elaborate mechanisms, including the so-called ' electronic brain,' " which have shown an uncanny ability to simulate human behaviour, and thereby to throw light on its possible nature."

Descartes compared a living animal to an automaton ; the kind of machinery he had in mind was that of the clockwork doll, jerking out the same old dance in time with the same old tune. Wiener prefers to compare it to a high-speed electrical computing machine, " the most complicated type of machine yet made," which receives through its ' sense organs ' a varying ' input ' from the outer world, and thereupon eventually produces an appropriately varying ' output,' regulated and controlled by information received, which can store up messages, exercise judgement and selection, and even pass on the information thus collected, in an elaborately analysed or synthesized shape, to some other machine.

[1] Reviewed by Mr. Donald Sholl in this *Journal*, II, pp. 122–4.

[2] It would seem that very much the same ideas occurred to a number of different workers at about the same time. One of the first to study the logical possibilities of modern machines appears to have been a British worker, D. S. Turing (of the National Physical Laboratory). Shannon, before joining the Bell Telephone Company, took his doctorate at the Massachusetts Institute of Technology with a thesis on the application of Boole's theory of classes to the study of systems of switching in electrical engineering, and has since deduced (by somewhat different methods) very similar formulæ. Wiener, in deriving his equations, states that he has used a personal communication from von Neumann.

"The very function of consciousness," wrote Fichte a century ago, "is to collect and use information." Similarly, Wiener, in his theories of mechanical and neural communication, insists that information rather than energy must form the central concept. And the fundamental task for those who design these "masterpieces of modern engineering" is, he contends, "to minimize the wasteful loss, not of energy, but of information."

Now to do this it is necessary to *measure information*. The term information may be defined, and its amount estimated, in a number of different ways. But the conception which Wiener develops is of particular interest to the psychologist, because it treats communicable information as consisting essentially in an orderly structure, or, as the psychologist might be tempted to say, a *Gestalt*. Hence, for Wiener, "the problem of measuring the amount of information is of a piece with that of measuring the regularity and irregularity of a pattern . . . of finding a measure of order, or of its negative—disorder." This is by no means a new idea in mechanics. Boltzmann, it may be remembered, in seeking to reinterpret thermodynamics in terms of statistical concepts, had already described 'entropy' as 'a measure of disorder' (i.e., of randomness), and observed that it appeared closely related to 'missing information,' since it depends on "the number of alternatives which remain possible to a physical system after all the information that is macroscopically observable has been allowed for." Von Neumann has treated information in sub-macroscopic physics along very similar lines: (*Mathematical Foundations of Quantum Mechanics*, 1932, chap. V).

In an appendix to his first chapter Wiener attempts, by a verbal rather than a mathematical argument, to make his own proposed method of measurement at once intelligible and plausible. "Because order is essentially lack of randomness," he writes, "the *more probable* a particular configuration, simply as a result of the play of chance, the *less order* it will contain. Hence the measure of the degree of order of a set of patterns, selected from a much larger set, is the negative of the logarithm of the probability of the smaller set, when the probability of the larger set is taken to be one." On this view information is considered to be essentially *selective*. To put the matter a little paradoxically, the information communicated by a given message depends not on what it says but on all the things it might have said and doesn't.

The psychological reader would probably find the argument easier to understand if it had been supplemented by a numerical example. Let us therefore make a very brief attempt. According to Wiener the amount of information communicated by a given message may be regarded as equal to the *number of correct decisions*, *n* say, which the receiver of the message would have to make between equally probable alternatives in order to determine which of the *N possible* messages was in fact *actually* sent. By way of illustration, let us suppose the information concerns the rating of some particular child for some psychological trait, and that the rating will state his position on a scale of eight indivisible intervals. *A priori* (we may suppose) any one of the eight figures is equally likely, so that the distribution is rectangular, as in assigning ranks. Now each successive decision can best be expressed by one of a succession of factors; and, if every decision turns merely on a simple disjunction, the factor-measurements need differ only in sign. Thus we begin by asking: is the child above or below the mean? Here are two equally probable alternatives. To decide this first question one factor will be sufficient; a plus sign can be taken to mean 'yes,' and a minus sign 'no.' If the verdict is that he falls (say) in the lower *half*, then a second factor will be needed to decide in which *quarter* of the total range he falls. And finally a third factor will be needed to place him in the appropriate *eighth*. Thus, with a series of three uncorrelated, undifferentiated, bipolar factors, we can reach any one of eight measurements, as shown below:

TABLE I. DECISIONS EXPRESSED AS FACTORS

Scale Measurement	Deviate	Weighted Factor-Measurements I	II	III	Sum
0 — 1	$-3\frac{1}{2}$	$-\frac{1}{2} \times 4$	$-\frac{1}{2} \times 2$	$-\frac{1}{2} \times 1$	$-\frac{1}{2} \times 7$
1 — 2	$-2\frac{1}{2}$	$-\frac{1}{2} \times 4$	$-\frac{1}{2} \times 2$	$+\frac{1}{2} \times 1$	$-\frac{1}{2} \times 5$
2 — 3	$-1\frac{1}{2}$	$-\frac{1}{2} \times 4$	$+\frac{1}{2} \times 2$	$-\frac{1}{2} \times 1$	$-\frac{1}{2} \times 3$
3 — 4	$-\frac{1}{2}$	$-\frac{1}{2} \times 4$	$+\frac{1}{2} \times 2$	$+\frac{1}{2} \times 1$	$-\frac{1}{2} \times 1$
4 — 5	$+\frac{1}{2}$	$+\frac{1}{2} \times 4$	$-\frac{1}{2} \times 2$	$-\frac{1}{2} \times 1$	$+\frac{1}{2} \times 1$
5 — 6	$+1\frac{1}{2}$	$+\frac{1}{2} \times 4$	$-\frac{1}{2} \times 2$	$+\frac{1}{2} \times 1$	$+\frac{1}{2} \times 3$
6 — 7	$+2\frac{1}{2}$	$+\frac{1}{2} \times 4$	$+\frac{1}{2} \times 2$	$-\frac{1}{2} \times 1$	$+\frac{1}{2} \times 5$
7 — 8	$+3\frac{1}{2}$	$+\frac{1}{2} \times 4$	$+\frac{1}{2} \times 2$	$+\frac{1}{2} \times 1$	$+\frac{1}{2} \times 7$
Square Sum	42	+32	+8	+2	$\frac{1}{4} \times 168$

The weights are, for factor I, 2^2; factor II, 2^1; factor III, 2^0; and similarly with a larger number of factors. We may verify the correctness of the analysis by noting that the variance due to the separate factors sums to the total variance, and this can be directly calculated; it is $\sigma^2 = \frac{1}{12}(N-1)N(N+1)$, that is, in the present instance, $\frac{1}{12} \times 7.8.9 = 42$.

What then is the relation between the number of decisions and the number of possible measurements? If the number of factors is n, the number of possible measurements will be $N = 2^n$; conversely, with N possible measurements in the scale, the number of factors will be $n = \log_2 N$. And a similar logarithmic expression can be used to state the amount of information supplied by each of the factors separately. Thus, if there are N alternatives, each equally likely, the probability of *each* will be $p_i = 1/N$; so that $H_i = -\log_2 p_i = \log_2 N$. Let us add the information supplied by each in turn. The total possible range is $(N - 0)$ units; and the range within which the required measurements are judged to fall after any one decision can be expressed as $(N_2 - N_1)$ units. Accordingly, the information supplied by such a decision will be, not the ratio of the two ranges, but the negative of its logarithm (negative, because otherwise the logarithm of the fraction would itself remain negative), i.e., $H = -\log_2 \frac{N_2 - N_1}{N - 0}$. With the first factor it will be $H_1 = -\log_2 \frac{4 - 0}{8 - 0}$ $= -\log_2 \frac{8 - 4}{8 - 0} = \log_2 2 = 1$; with the first and second factors together it will be $H_{1+2} = -\log_2 \frac{2 - 0}{8 - 0} = \log_2 4 = 2$; and with all three factors it will be $H_{1+2+3} = \log_2 8 = 3$. If the alternatives are not equally likely, they must be weighted; and we obtain $H = -\Sigma p_i \log_2 p_i$. From the expression thus reached, it will be seen, we can increase information either (*a*) by increasing the number of possible choices or (*b*) by making the probabilities of the various choices as nearly equal as circumstances permit.

If the distribution is not discrete but continuous, we can apply the same principles. Let us suppose that the *a priori* probability that a certain quantity lies between x and $x + dx$ is $p_1(x)dx$, and that, after our observations have been made, the *a posteriori* probability becomes $p_2(x)dx$: what (asks Wiener) is the amount of new information added by the determination of this *a posteriori* probability? The answer is given by the continuous form of the expression already reached in the discrete case, namely, $H = -\int_{-\infty}^{\infty} p(x) \log_2 p(x)dx$. This expression is similar to that derived by Boltzmann in his attempt to relate entropy to probability. Hence the adoption of Boltzmann's symbol H.

It is natural to enquire how Wiener's formula differs from Fisher's. The difference, I fancy, can be most simply exhibited by considering how it applies when the distributions are normal. To begin with, it is easy to show that, in both the one-dimensional and the multi-dimensional cases, the distribution function which gives a maximum entropy is that of the normal curve. With such a distribution we then have, for the one-dimensional case, $p(x) = \frac{1}{\sqrt{2\pi\sigma}} exp - (x^2/2\sigma^2)$; and, on substituting this in the expression just given for H, that at once reduces to $H = \log (2\pi e)^{\frac{1}{2}} \sigma$. For the n-dimensional case we obtain $H = \log (2\pi e)^{n/2} \left| a_{ij} \right|^{-\frac{1}{2}}$, where $\left| a_{ij} \right|$, as usual, denotes the determinant of the associated quadratic form. Now with Fisher information is defined to mean "that quantity which characterizes a body of data with respect to a parameter, regardless of the method used to estimate that parameter"; and the amount of information communicated by any efficient statistic is measured by "the reciprocal of the variance of that statistic." Comparing the two expressions we see that, in both cases, information is increased by narrowing the range within which the measurement falls; but Wiener secures an additive measure by taking the *logarithm* of the measure of that range, i.e., of the standard deviation ($\log k\sigma$), whereas Fisher does so by taking it *square* (σ^2).

Let us now introduce a 'specific factor' to represent the effect of randomness or error—what Wiener terms, with the analogy of telephonic communication in mind, 'noise.' Let the 'true message' be denoted by g, the obscuring 'noise' by s, and the message as received (i.e., when more or less contaminated with 'noise') by $x = g + s$; let the variance of x be $\sigma_x^2 = \sigma_g^2 + \sigma_s^2$; and let the probability functions for g and s be normal. Then, as Wiener has formally proved (with a rather different notation), the information gained, when x is fixed by observation, will be $\frac{1}{2} \log_2 \frac{\sigma_g^2 + \sigma_s^2}{\sigma_s^2}$. Thus, when the 'noise' is *nil*, the information becomes infinite; and, as the 'noise' increases indefinitely as compared with the 'true message,' the information sinks to zero. Since two 'factors' only have so far been assumed, the resulting formula corresponds with what might be called a 'Spearman analysis' of the situation. But, of course, in actual practice both variables and factors are bound to be multiple; and, in his earlier volume, Wiener has further shown how the simple formula can be extended and generalized to meet such a case. This, as he has told us, constitutes his 'formal solution' of what he terms "the message-noise problem and the central problem of communication engineering."

The greater part of his new book consists in an attempt to describe, in non-technical language, what he takes to be the implications of the foregoing mode of approach, when applied to the inter-

pretation of neurological, psychological, and current sociological problems. Having explained in his first two chapters his conception of communication and of information, as developed in modern engineering, Wiener turns next to the consideration of " the rather special sort of machine known as a human being." The central nervous system, or what it is ordinarily called the individual mind, may, he suggests, be usefully regarded as a " terminal machine in a communication network." Any such mechanistic standpoint is, as he realizes, liable to criticism on two main grounds : first, that mental activity exhibits purposive control, and secondly, that it exhibits the phenomena of learning. Wiener's answer is that both these characteristics can be found in modern machines, and arise from the use of ' feed-back.' " The feed-back principle means that behaviour is scanned for its result, and that the success or failure of this result modifies future behaviour." And " learning is a particularly complicated form of feed-back." He claims that it is quite possible to give " a method of constructing a machine that can learn." His proposal requires us to think of the brain " as an analogy machine " rather than as a " digital machine " ; [1] this would enable us to exploit the possibilities of " an undirected message, spreading out until it finds a receiver, which is then stimulated by it." In short, according to his theory (though he does not himself express it in that way), both purpose and learning consist essentially in processes whereby a complex self-adjusting system, once disturbed, tends to seek a new position of equilibrium.

As Wiener acknowledges, the idea and even the design of a ' mechanical brain ' is by no means a novelty. It is at least as old as the ' analytical engines ' planned, and partly constructed, by Charles Babbage, who was Professor of Mathematics at Cambridge from 1828 to 1839. And the materialists of the nineteenth century—those who were not engineers perhaps more often than those who were—claimed that here was formal proof that La Mettrie's conception of *l'homme machine* would soon be realized in the concrete.[2] Probably most psychologists, at any rate in this country, will agree with the reply made by Lady Lovelace, who pointed out that, if machines like those of Babbage appeared to carry out purposive operations, that was merely because those purposes were first conceived by the designer : " the analytical engine," she writes, " has no pretensions whatever to *originate* anything : it can do only what we *order* it to perform." [3]

Wiener himself repeatedly emphasizes that, no matter how much a human being may resemble a machine, it is a gross mistake to *treat* him simply as a machine. Indeed, his great fear is that our growing knowledge of the processes of communication may lead " those who suffer from a power complex," whether " strong men in business " or " equally ruthless totalitarian governments," to turn individuals into mere wheels or cogs in the vast political machine. Accordingly, the greater part of his book is concerned with a miscellaneous set of sociological corollaries which he thinks may be drawn from the general principles he has laid down. In most cases, however, the arguments adduced are somewhat vague : and the reader is left with the feeling that many of Wiener's political and ethical views, though expressed in his own special terminology, have been formed on quite different grounds.

He begins with language as the basis of all other social activities. Man owes his eminence over the lower creation to his gift of speech ; and speech is pre-eminently " a mode of communication." It rests, he believes, on a special kind of ' built-in mechanism,' namely, an " innate interest in coding and decoding," a mechanism which human beings inherit and chimpanzees do not. He briefly discusses in turn the ' phonetic aspect,' the ' semantic aspect,' and what he calls ' behaviour language.' In each case, he tells us, by using the formulæ already reached, we could evaluate the actual and the possible amount of information conveyed ; and at each stage we should find, in accordance with the ' cybernetic form of the second law of thermodynamics,' that information, like available energy, is

[1] The typical ' digital machine ' carries out its numerical operations by counting discrete objects (e.g., a desk computer such as the Marchant which counts the cogs of its gear wheels) ; the ' analogy machine ' translates numbers into physical quantities, of which those numbers are the measures. and operates with these quantities (e.g., the common slide rule, the integraph, and the differential analyser).
[2] Like its modern counterpart, and unlike the other computing machines envisaged by Babbage, his ' analytical engine ' had both a ' store ' or memory (which accomplished something very like learning and retention) and a ' control ' (which accomplished something very like purposive direction). It was these, rather than the ' mill,' which only carried out ' the mere logical operations of computing,' that suggested a comparison with the human mind. It should perhaps be explained that our most detailed knowledge of these machines is derived from the discussion with Babbage recorded by cf. General Menebrea, which were translated by the Countess of Lovelace (daughter of the poet Byron) : Menebrea, L. F., *Scientific Memoirs* (Eng. trans., 1842). Lady Lovelace herself had apparently given Babbage some mathematical assistance : for instance, he quotes her as correcting one of his recurrent formulæ.
[3] It is instructive to note that, more recently, no less an authority than Professor D. R. Hartree has protested, on much the same grounds, against the employment of phrases like ' mechanical brain ' and ' electronic brain,' and the inferences popularly drawn from them : as he says, all the purposive thinking that such machines display has really been " done beforehand by the designer " (*The Times*, Nov. 7, 1946).

always lost, never gained. However, he gives no concrete illustrations to show in any detail how his principles could be applied in psychological researches on such problems.

By way of an example, we may perhaps take a familiar psychological problem from the ' semantic aspect.' How can we determine (i) the average ability and (ii) the individual abilities of children of a particular age to describe their experiences in words ? Enquiries show that the average London child of ten has a vocabulary of about 6,000 words, and speaks in sentences averaging about eight words in length. By the formula given we find that, if he were to select each word independently, the number of events he could describe would amount to $6{,}000^8$ or about 10^{30}. Further, if, with the communication engineers, we adopt the binary unit or ' bit ' as our unit, and assume that the child speaks at a rate of two words per second, we shall find that, measured as Wiener proposes, the amount of information which he can thus convey per second is $2 \log_2 6{,}000 = 25$ bits (approx.). But, of course, the probabilities for the later words in any sentence are by no means independent of the preceding. Making due allowance for this dependence, and applying the appropriate theorems from the theory of communication as expounded by Wiener and Shannon, I calculate that the number of events the child could describe would be about 10^{12}, and the information he could convey per second would be about 20 bits. In like manner, we could, if we wished, compare the relative merits of the vocabularies of (say) basic English and the ordinary educated Englishman, or again of the latter and a linguistic genius like Shakespeare. Here plainly is a new and suggestive approach for investigating those innumerable questions in the psychology of language which literary critics like Vernon Lee and statisticians like Udny Yule have each attempted to study by his or her own quantitative procedure.

The omission of vowels in Hebrew and Pitman's shorthand shows very strikingly how much can be left to the probabilities developed in the mind of the reader. Pitman's use of hooks, circles, grammalogues, stereotyped contractions, and the like, indicates that he also relied on what Shannon calls ' transitional probabilities.' Indeed, we are told that, before completing his system, he made a rough survey to ascertain the frequencies of such recurrent sequences.

The possibilities of the method in dealing with problems from the ' phonetic aspect ' are perhaps not quite so obvious, but they come nearer to the everyday questions with which the modern communication engineer is concerned. Those who attend annual conferences must be only too well aware that, in a large and noisy audience-hall, some speakers are quite inaudible, while others can be heard as easily as a trained actor : in listening to opera, a libretto sung in German is generally easier to follow than one sung in Italian or French, and English is the obscurest of all. Here, plainly, both the speaker (or the singer) and his hearers form part of a communication system, as defined by Wiener ; and the problems raised therefore should be susceptible of treatment along the lines he describes. What, for example, are the special psychological conditions governing the efficient utterance and the efficient reception of articulate speech ? If we may accept Professor Daniel Jones's analysis,[1] the phonemes of standard English speech comprise 13 vowels and 23 consonants, which are uttered in ordinary conversation at an average rate of about 12 per second : other European languages do not diverge very widely. The relative frequencies of these constituents and of their possible combinations have been ascertained.[2] It should therefore be quite easy to calculate the actual and possible information conveyed by such a system, to study individual differences, and to compare the effects of various factors—speed, loudness, background noise, reverberation, and the like. Something of this sort has already been undertaken for military purposes during the war : for example, in a study of spoken messages, partly masked by aeroplane noise, it is found that loudness of sound and precision of consonants are of far greater influence than rate of speaking, and that (owing no doubt to the fact that we perceive relations between sounds rather than absolute sounds) the human recipient is far less efficient than a mechanical recipient, but on the other hand is far less disturbed by distortion.[3]

From communication between one individual and another by audible language and visible behaviour Wiener passes to a discussion of communication with groups. In countless ways, so he contends, modern mechanical techniques have vastly enhanced our powers of communication, and thus rapidly extended the social possibilities, whether for good or for evil, of man as organized in society. " Literature and the arts," we are told, " have already become increasingly dominated by mass communication " ; and the issues of politics and war are determined largely by devices of communication, both true and false. Following a hint thrown out by von Neumann and Morgenstern in their *Theory of Games*, Shannon has quite recently suggested that, in the near future, it may be quite practicable to construct a mathematical machine " to evaluate military situations and determine the best move at any specific stage." Père Dubarle has even envisaged a colossal ' *machine à gouverner*.' Such proposals, so Wiener believes, are " no longer moonshine, but imminent and

[1] Jones, D., *An Outline of English Phonetics* (1939) : cf. Jespersen, O., *Lehrbuch der Phonetik*, 1904.

[2] French, N. R., *et al.*, ' The Words and Sounds of Telephone Conversations,' *Bell System. Techn. Journ.*, IX, 1930, pp. 290–324.

[3] Cf. Abrams, M. H., *et al.*, ' Speech in Noise : A Study of Factors Determining its Intelligibility ' (1944), and other papers of the Harvard Psycho-Acoustic Laboratory ; Black, J. W., and Mason, H. M., ' Training for Voice Communication,' *J. Acoust. Soc. Amer.*, XVIII, 1946, pp. 441–5.

dangerous contingencies." We are thus, he thinks, witnessing a "second industrial revolution," which threatens to be even more revolutionary than the first.

Nevertheless, I fancy that the majority of his readers will feel that these more sweeping and somewhat trite conclusions, which he has deduced about modern civilization and its current trends, are less informative than the incidental corollaries and analogies which he has thrown out for the study of individuals. Some of his most interesting inferences arise from his endeavours to interpret nervous disorders by comparing them with somewhat similar phenomena that may be observed when a computing machine gets out of order. Thus, to illustrate the implications of his theory he describes small run-about machines "with a simple built-in purpose"—the 'Moth' (phototropic) and the 'Bedbug' (photophobic)—not unlike those which readers in this country may have seen, designed by Dr. Grey Walter of the Burden Neurological Institute. Wiener explains how, under certain conditions such as overloading, a kind of conflict may arise between two quasi-postural motors, incorporated in the machine, with the result that its rudder goes into a sort of tremor, resembling (so he suggests) the shaking palsy of Parkinsonism, or makes coarser oscillatory movements, like the indecisive wavering of the anxiety-neurotic. His comparisons will not perhaps seem very convincing to the neurologist. But a psychologist may be permitted to wonder whether his formulæ might not be usefully extended to the experimental study of oscillation in normal rather than in abnormal processes. Dr. Philpott, for example, has recently suggested that the well-known fluctuations in mental work may be compounded out of oscillations having the periodic rhythm of the cyclic processes of the electron. As I have indicated elsewhere, it would seem more likely that they are related to the rhythmical electro-chemical changes characteristic of nerve-cell processes. Adopting Wiener's terminology, we may regard our whole nervous equipment as an intricate communication system in which the signals are transient pulses, having a relatively constant magnitude. And it is now known that the response of a nerve cell to a single stimulation consists, not of a single pulse or wave, but of a train of oscillating changes. Work on nerve conduction in the lower animals, and still more recently on the response of receptor organs in man (the Pacinian corpuscles, for example), has shown that the result is a series of discharges, diminishing logarithmically (or rather exponentially) in strength, and spaced at fairly regular intervals. The rate of discharge, at least for the most typical cases, has been quite accurately ascertained. Assuming that a periodicity of this kind is characteristic of all neural processes, we have a possible basis for explaining the rhythm and the course of observable mental fluctuations. Or again the oscillations may be related to the rhythmic sequences by means of which the apperceptive process analyses a complex object. In experiments on attention to visual schemes ('spot patterns' and the like) Dr. Flugel and I noted how the 'fluctuating' type of mind ranges to and fro across the pattern, much as in reading. This suggests an analogy rather with the way 'information' is serialized by 'scanning' for purposes of communication by television, except that the rhythmic sequences die away exponentially.[1] Here, then, are three alternative hypotheses. By the aid of mathematical deductions similar to those developed by Wiener, we could derive distinctive corollaries from each one, and then put them to a crucial experimental test.

To my mind it is in these humbler, incidental applications, which the book so constantly suggests to the mind of a psychological reader, rather than in the vaguer if profounder reflections regarding world affairs that the interest of the book really lies. Within the field of physical engineering, as will be gathered from the instances cited above, Wiener is always ready to use his mathematical ingenuity to derive new notions, which he then objectively checks by constructing an actual model to embody them. Yet, when he turns to the problems of psychology, he seems content with mere speculation and analogy. After all, in an empirical science, the chief value of such mathematical techniques resides in their power to deduce precise conclusions for experimental testing and to predict fresh results which can be confirmed or confuted by systematic observation. How fruitful for the psychological experimenter the ideas and hypotheses derived from modern engineering devices may be has been abundantly shown by Kenneth Craik's description of the human operator as an 'engineering system.'[2] Nevertheless, Wiener himself, in his psychological and sociological chapters, has apparently few concrete suggestions to make for future research: dogmatic deductions take the place of inductive proofs or plans for experimental enquiries. But perhaps Wiener would say that his own task has been to hammer out a serviceable tool, and that its application must be left to the psychologists themselves.

C. B.

[1] Long ago Stout remarked how, "in looking hard at a man, we *scan* his proportions, by successive eye-movements with corresponding perceptions" (*Analytic Psychology*, II, p. 124). Dr. Grey Walter suggests, from a comparison of the rates, that the alpha rhythm of the visual cortex may be related to these scanning processes.

[2] Cf. Craik, K. J. W., 'Theory of the Human Operator in Control Systems.' *Brit. J. Psych., Gen. Sect.*, XXXVIII, 1947, pp. 56–61.

INDEX TO VOLUME IV

INDEX OF AUTHORS

Authors	*Titles*	*Page*
Banks, C., and Sinha, U.	An Item-Analysis of the Progressive Matrices Test	91–94
Barakat, M. K.	A Factorial Study of Mathematical Abilities	137–156
Bartlett, M. S.	A Further Note on Tests of Significance in Factor Analysis	1–2
Burt, Cyril	The Numerical Solution of Linear Equations	31–55
Burt, Cyril	Test-Construction and the Scaling of Items	95–129
Burt, Cyril, and Watson, H.	Factor Analysis of Assessments for a Single Person	179–191
Cattell, R. B.	A Factorization of Tests of Personality Source Traits	165–178
Durbin, J.	Incomplete Blocks in Ranking Experiments	85–90
Ehrenberg, A. S. C.	Note on the Normal Transformation of Ranks	133–134
Emmett, W. G.	The Estimation of Correlation Coefficients from the Variance of Score Differences	55–60
Gourlay, N.	Difficulty Factors Arising from the Use of Tetrachoric Correlations in Factor Analysis	65–76
Hick, W. E.	Information Theory and Intelligence Tests	157–164
Jones, W. R., and Stewart, W. A. C.	Bilingualism and Verbal Intelligence	3–8
Peggs, A. D.	An Analysis of Variance with Unequal Numbers	77–84
Sinha, U., and Banks, C.	An Item-Analysis of the Progressive Matrices Test	91–94
Slater, P.	The Transformation of a Matrix of Negative Correlations	9–20
Stewart, W. A. C., and Jones, W. R.	Bilingualism and Verbal Intelligence	3–8
Sutherland, J.	Factor Rotation by the Method of Extended Vectors	21–30

INDEX OF SUBJECTS

Titles	*Authors*	*Page*
Analysis of Variance with Unequal Numbers, an	Peggs, A. D.	77–84
Bilingualism and Verbal Intelligence	Jones, W. R., and Stewart, W. A. C.	3–8
Correlation Coefficients, the Estimation of, from the Variance of Score Differences	Emmett, W. G.	55–60
Difficulty Factors Arising from the Use of Tetrachoric Correlations in Factor Analysis	Gourlay, N.	65–76
Factor Analysis, A Further Note on Tests of Significance in	Bartlett, M. S.	1–2
Factor Analysis of Assessments for a Single Person	Burt, Cyril, and Watson, H.	179–191
Factor Rotation by the Method of Extended Vectors	Sutherland, J.	21–30
Information Theory and Intelligence Tests	Hick, W. E.	157–164
Linear Equations, the Numerical Solution of	Burt, Cyril	31–55
Mathematical Abilities, a Factorial Study of	Barakat, M. K.	137–156
Negative Correlations, the Transformation of a Matrix of	Slater, P.	9–20
Personality Source Traits, a Factorization of Tests of	Cattell, R. B.	165–178
Progressive Matrices Test, an Item-Analysis of the	Banks, C., and Sinha, U.	91–94
Ranking Experiments, Incomplete Blocks in	Durbin, J.	85–90
Ranks, Note on the Normal Transformation of	Ehrenberg, A. S. C.	133–134
Test-Construction and the Scaling of Items	Burt, Cyril	95–129

BOOKS REVIEWED

Authors	*Titles*	*Page*
Blackwell, A. M.	A List of Researches in Education and Educational Psychology	136
Johnson, P. O.	Statistical Methods in Research	62–63
Scottish Council for Research in Education	Traditional Number Rhymes and Games	64
Scottish Council for Research in Education	Studies in Reading	135
Shannon, C. E., and Weaver, W.	The Mathematical Theory of Communication	193–195
Vernon, P. E.	The Structure of Human Ability	61–62
Wiener, N.	The Human Use of Human Beings	195–200

BIOL. LONDON. UNIV.

UNIVERSITY OF LONDON LIBRARY
DO NOT REMOVE

Vol. IV Part I March 1951

THE BRITISH JOURNAL OF PSYCHOLOGY

Statistical Section

EDITED BY
CYRIL BURT AND GODFREY THOMSON

WITH THE ASSISTANCE OF
CHARLOTTE BANKS AND ARTHUR SUMMERFIELD
AND THE FOLLOWING EDITORIAL BOARD

A. C. AITKEN
M. S. BARTLETT
W. G. EMMETT
M. G. KENDALL
D. N. LAWLEY
E. S. PEARSON
L. S. PENROSE
J. FRASER ROBERTS
A. RODGER
B. BABINGTON SMITH
W. STEPHENSON
P. E. VERNON

Managing Editor: H. G. MAULE

UNIVERSITY OF LONDON PRESS LTD.
WARWICK SQUARE, LONDON, E.C.4.

Price Fifteen Shillings

CONTENTS

		Page
BARTLETT, M. S.	A Further Note on Tests of Significance in Factor Analysis	1
JONES, W. R., and STEWART, W. A. C.	Bilingualism and Verbal Intelligence . . .	3
SLATER, P.	The Transformation of a Matrix of Negative Correlations	9
SUTHERLAND, J.	Factor Rotation by the Method of Extended Vectors	21
BURT, CYRIL	The Numerical Solution of Linear Equations. .	31
EMMETT, W. G.	The Estimation of Correlation Coefficients from the Variance of Score Differences	55
BOOK REVIEWS		61

The British Journal of Psychology (*Statistical Section*) is issued by the British Psychological Society. Three numbers will be published annually.

The subscription price is 30*s.* a volume of about 200 pages. Subscriptions should be sent to any Bookseller, or to THE MANAGER, University of London Press, Ltd., Warwick Square, London, E.C.4.

Papers for publication should be sent either to Professor Sir CYRIL BURT, Psychological Department, University College, London, W.C.1, or to Professor Sir GODFREY THOMSON, Moray House, Edinburgh, 8. Contributors will receive 25 copies of their papers free : extra copies may be had at cost price, if the order is given when the proofs are returned.

Authors are asked to send their papers in a form which is ready for submission to the printer : that is to say, the arrangement of the manuscript (including a brief list of contents at the head and a summary of results and conclusions at the end) should follow the general arrangement observed in the publications of the British Psychological Society.

Books for review should be sent to one of the Editors ; advertisements for insertion to the Manager, University of London Press, Ltd.

The British Psychological Society also issues quarterly *The British Journal of Psychology, General Section*, and *The British Journal of Medical Psychology*. Subscriptions, 40*s.* per annum for either journal, should be sent to the Cambridge University Press, Bentley House, Euston Road, London, N.W.1. The Society also publishes, jointly with the Training College Association, *The British Journal of Educational Psychology*. Subscriptions, £1 per annum, for this Journal should be sent to Methuen & Co., Ltd., 36 Essex Street, London, W.C.2.

Papers for publication in the *General Section* should be sent to Professor D. W. HARDING, the Psychological Laboratory, Bedford College, London, N.W.1 ; those for publication in *The British Journal of Medical Psychology* should be sent to Dr. J. D. SUTHERLAND, 2 Beaumont Street, London, W.1 ; those for publication in *The British Journal of Educational Psychology* to Emeritus Professor C. W. VALENTINE, The White House, Wythall, Nr. Birmingham.

BRITISH JOURNAL OF EDUCATIONAL PSYCHOLOGY

Edited by C. W. VALENTINE

Vol. XXI. Part 1.

JAMES DREVER (Jnr.). James Drever (1873-1950).
JOHN SUTHERLAND. A Comparison of Pupils' Arithmetical Ability in the Secondary School with their Ability at the Time of their Transfer from Primary Schools.
HILDE T. HIMMELWEIT & ASENATH PETRIE. The Measurement of Personality in Children.
E. A. PEEL & D. RUTTER. The Predictive Value of the Entrance Examination as Judged by the School Certificate Examination.
D. G. BURNS. An Investigation into the Extent of the First-year Vocabulary in French in Boys' Grammar Schools.
S. M. BENNETT. A Group Test of Hearing for Six-year-old Children.
HANNAH STEINBERG. List of Publications by Sir Cyril Burt.
SIR CYRIL BURT. Critical Notice : The Structure of Human Abilities.

The Journal *appears once each University term—in February, June and November*

Subscription : Twenty shillings per annum post free, payable in advance, or 10*s.* per number

Subscriptions should be sent direct to the Manager, B.J.E.P. Department

MESSRS. METHUEN & CO., LTD

36 ESSEX STREET, STRAND, LONDON, W.C.2

Members of the British Psychological Society receive the *Journal* on special terms

HUMAN RELATIONS

A Journal of Studies towards the Integration of the Social Sciences

VOL. IV, No. 1

E. L. TRIST and K. W. BAMFORTH. Some Social and Psychological Consequences of the Longwall Method of Coal-getting.

HAROLD H. KELLEY. Communications in Experimentally Created Hierarchies.

MILTON W. HOROWITZ, JOSEPH LYONS, and HOWARD V. PERLMUTTER. Induction of Forces in Discussion Groups.

LILY HERBERT. A Case Study of a Group Selection Procedure.

P. BLACK. Toward a Systematic Field of Race Relations.

30*s. per annum or* **8***s.* **6***d. per issue*

The Tavistock Institute of Human Relations, London, England. The Research Center for Group Dynamics, Ann Arbor, Mich.

Subscriptions and orders to Managing Editor

TAVISTOCK PUBLICATIONS LTD

2, BEAUMONT STREET LONDON, W.1

THE BRITISH JOURNAL OF PSYCHOLOGY (GENERAL SECTION)

Edited by D. W. HARDING

Vol. XLI Parts 3 & 4 **December 1950** **12s. 6d. net.**

RUSSELL, ROGER W. The comparative study of 'conflict' and 'experimental neurosis'.

HALL, K. R. L. The effect of names and titles upon the serial reproduction of pictorial and verbal material.

BARTLETT, F. C. Incentives.

PRASAD, JAMUNA. A comparative study of rumours and reports in earthquakes.

JAMES, H. E. O., and TENEN, C. How adolescents think of peoples. I.

JACKSON, LYDIA. Emotional attitudes towards the family of normal, neurotic and delinquent children. Part II.

FITT, A. B., and ROGERS, C. A. The sex factor in the Cattell intelligence tests, Scale III.

PUBLICATIONS RECENTLY RECEIVED.

NOTICE.

The subscription price per volume, payable in advance, is 40s. net (post free).

Subscriptions may be sent to any bookseller or to the

CAMBRIDGE UNIVERSITY PRESS

BENTLEY HOUSE, EUSTON ROAD, LONDON, N.W.1

OCCUPATIONAL PSYCHOLOGY

January, 1951

CONTENTS

An Examination of the Notion of Communication in Industry. By P. H. Cook.

Socio-Psychological Factors in Productivity. By R. Marriott.

Levels of Expectation in Productivity. By R. G. Stansfield.

Twenty-Five Years of Research in Vocational Guidance. By Percival Smith.

Current Research Needs in Vocational Guidance. By Alec Rodger.

Absenteeism, Turnover and Morale in a Metals Fabrication Factory. By Willard A. Kerr, George J. Koppelmeier and James J. Sullivan.

The Vocational Guidance and Employment of Handicapped and Disabled Young People. By William Duncan.

The Technique of Role-Reversal : A Preliminary Note. By John Cohen.

Human Ability : A Supplementary Note.

Book Reviews.

Annual Subscription £1

National Institute of Industrial Psychology

14, Welbeck Street, London, W.1

UNIVERSITY OF LONDON
13 JUN 1951
LIBRARY

Printed in England by Hazell, Watson & Viney, Ltd., London and Aylesbury—12630-51